# 大規模相互作用ネットワーク解析手法の開発

博士学位論文

# 大規模相互作用ネットワーク解析手法の開発

Development of a method for analysis of large interaction networks

東北大学大学院情報科学研究科

応用情報科学専攻

生命システム情報科学分野

田高 周

2017 年 2 月

# 目次

目次

# 図表目次

**目次**

**目次**

目次

# 要旨

近年、生命科学分野においてシステム生物学という考え方が市民権を得るようになった。システム生物学とは、生物をある一つのシステムとして考え、そこにシステム工学の考え方や解析方法を導入することで生命を理解しようとする学問である。近年では、実験技術などの向上により、DNA 配列や mRNA 発現データ・タンパク質間の相互作用・メタボローム解析データなど多種多様な情報が、網羅的、かつ、高速に生産されるようになった。そして、細胞内外における分子ネットワークに関するデータと知見が蓄積されるようになった。その結果として、生物の中で部品（分子）がどのように連携し、高度なシステムとして働くのかという問いに対し、一定の答えを返せるようになってきた。

生命システムは、細胞内外の分子の相互作用によって複雑なシステムとして働く。分子間の相互作用は情報科学分野ではネットワークとしてモデル化可能である。従って、細胞内の分子ネットワークの解析を行うことで生命現象の謎の解明に挑戦することができると期待される。本論文の第 1 章ではタンパク質間相互作用ネットワークから機能モジュールを抽出する手法の開発を行う。様々な生命現象はタンパク質の相互作用の結果として引き起こされるものであり、ある特定の生命機能を持つタンパク質は、互いに相互作用することが期待される。そのため、互いに相互作用するタンパク質群はタンパク質間相互作用ネットワーク上では密に接続されたクラスタを形成していると考えられる。逆に考えると、タンパク質間相互作用からクラスタを抽出することで、機能モジュールと呼ばれる、互いに関連したタンパク質の群を抽出することが可能であると考えられる。機能モジュールは生命システムの基本的なコンポーネントである。生命システム全体のダイナミクスを検討する上で、機能モジュールの同定は大変重要である。

本論文の第 2 章では複数の分子ネットワークの統合を試みる。これまでに生命を表現する分子ネットワークとしていくつかの種類のネットワークが考案されてきている。例えば、タンパク質間相互作用ネットワーク・遺伝子共発現ネットワーク・シグナル伝達ネットワークと呼ばれるものである。それぞれのネットワークを作成するために必要な実験や計算が異なり、また、異なった特徴を持っている。第 2 章では、タンパク質間相互作用ネットワークと遺伝子共発現ネットワークネットワークの統合を通じ、それぞれのネットワークの特徴づけを行い、また、統合による利点などを確認する。

第 3 章では、大規模生命関連データの解析補助を行うアプリケーションの開発に関して述べる。実験データの生産スループットの上昇は、生物に関する深い知見をもたらすが、同時に、生産されるデータの管理や解析方法といった点で問題を引き起こす。例えば、一度の実験で生産されるデータの量はテラバイトに及ぶこともあり、人の手で扱えるものではないという場合も多々発生している。そのような場において、生産されるデータを整理された形で蓄積し、さ

**要旨**

らに蓄積された膨大なデータの山から生物学的に意味のあることを見出すための基盤やアルゴリズムを開発することは情報科学分野として大変重要であり、同時に、意義のあることであると考えられる。

# 本研究への導入

**本研究への導入**

2000 年代初頭より、生命科学分野においてシステム生物学と呼ばれる新しい分野が認知されてきた。システム生物学は、生物学分野に工学的な考え方や解析手法を導入し、生物をある一つのシステムとして理解することを目標としている。このような概念は古くから存在する。その一つが、例えば 1940 年代に Norbert Wiener によって提唱されたサイバネティクス (Wiener, 1949) である。サイバネティクスは、一般に、生物と機械における通信・制御・情報処理を統一的に扱う学問・分野である。Wiener は、生物と機械では情報のやり取りや個体制御の行い方などの点で類似性があることに注目しサイバネィクスを提唱した。このような考え方はシステム生物学の先駆けであると考えることができる。しかしながら、サイバネティクスが発表された当初は、生体を構成する細胞や分子に関する知見が不足していたため、生物をブラックボックスとして扱い、現象論的な研究・解析しか行うことができなかった。

1930 年代に Warren Weaver により分子生物学と呼ばれる分野が提唱された。 (Weaver, 1970) 分子生物学は様々な生命現象を分子の動きや相互作用の結果として記述することを目的とした学問である。1953 年に James Watson、Francis Crick により DNA の二重らせん構造が発見された (Watson and Crick, 1953) ことを皮切りに分子生物学分野が急速に発展した。様々な実験手法が開発・改良され、様々な分子やその機能、また、分子の相互作用が発見された。その結果として、今日では、生物をシステムとして捉え、解析することが可能になってきた。

2002 年、Zoltan Oltvai、Albert-Laszlo Barabasi により、「Life's Complexity Pyramid」(Oltvai and Barabási, 2002) が発表された。

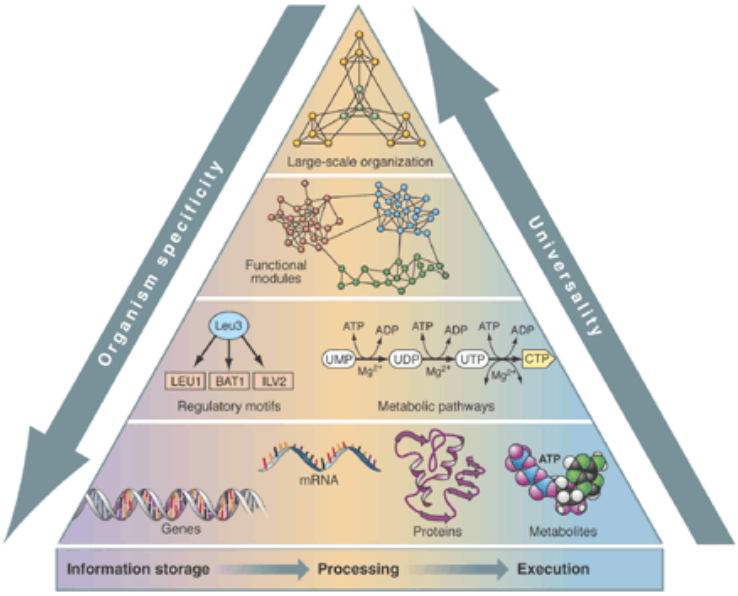



**図 1 Life's Complexity Pyramid (Oltvai and Barabasi, 2002)**

図 1 では、（i）DNA や mRNA、タンパク質などが互いに相互作用し、ある一つの代謝パスウェイを構成する（ii）代謝パスウェイ間も互いに関係しあい、さらに大きな機能モジュールを構成している（iii）機能モジュールが集合することによりさらに大きな一つのシステムがつくられるということが表現されている。システム生物学では、このように大きな一つの生命システムを小さいコンポーネントの相互作用で検討する。つまり、生物を分子間の相互作用を表すネットワーク・グラフとして表現することもできる。このような、遺伝子やタンパク質分子のネットワークに着目する考え方は、特に、「ネットワーク生物学」 (Barabási and Oltvai, 2004) と呼ばれる。

過去の研究によると (Girvan and Newman, 2002) 遺伝子ネットワークなどの相互作用ネットワークは、現実世界に存在する巨大で複雑なネットワークといくつかの共通点を持つことが知られている。例えば、スケールフリー性・スモールワールド性・クラスタ性である。従って、既存の情報科学的・工学的・数学的な手法を生命科学分野に持ち込むことで、複雑な生命システムネットワークの解析の足がかりになることが期待される。その点において、生命科学分野における情報科学の重要性は増していると考えられる。

その他にも生命科学分野において情報科学的アプローチが求められる点は存在する。例えば、情報爆発 (Latanya Sweeney, 2006; Datt, 2011) によるものが挙げられる。近年では、実験技術の向上・変化などの影響により、様々な種類の生命関連データの量が爆発的に増加している。例えば、タンパク質間の相互作用を検出するための実験手法として、TAP 法や共免疫沈降法などが用いられてきた。 (Berggård *et al.*, 2007; Rao *et al.*, 2014) これらの手法は、実験に対しかかる時間や費用などの点で大量のタンパク質間相互作用を調査するには難点があった。しかし、近年では、酵母ツーハイブリッド法やタンパク質マイクロアレイ、質量分析器を用いた手法など、様々な手法が開発されてきた。 (Rao *et al.*, 2014) 近年開発された手法では過去の手法と比べ、より多くのタンパク質間の相互作用をより簡便に同定することができるようになり、データ生産のスループットが上がった。また、実験的な手法の他に、計算機的にタンパク質間相互作用を予測する手法の開発も進んでいる 。(Rao *et al.*, 2014) タンパク質間相互作用データだけでなく、タンパク質立体構造データなど、様々かつ大量の生命関連データが利用可能になってきたためである。

データの生産は年々増加している。増え続ける生命関連データを蓄積するため、様々なデータベースが開発されている。 (Rigden *et al.*, 2016; Katayama *et al.*, 2013) このように増え続けるデータを整理された形で蓄積し、データベースとして提供するという点において情報科学の力が用いられている。

また、データを単に蓄積させていくだけでは意味がない。データの山から生命科学的に意味のある事象を発見することが重要である。増え続けるデータを人力でキュレーションすることは不可能に近いため、計算機による自動処理が求められる。この点においても情報科学が生命科学分野に介入することには意味があると考える。

# 1. 大規模相互作用ネットワーク中に含まれる機能モジュール解析手法の開発

# 1. 大規模相互作用ネットワーク中に含まれる機能モジュール解析手法の開発

「1. 大規模相互作用ネットワーク中に含まれる機能モジュール解析手法の開発」は以下の投稿論文を元にしている。

Shu Tadaka and Kengo Kinoshita, “NCMine: Core-peripheral based functional module detection using near-clique mining”, *Bioinformatics* (2016), doi: 10.1093/bioinformatics/btw488

# 1.1 導入

遺伝子共発現やタンパク質間相互作用（Protein-Protein interaction; PPI）などの、遺伝子やタンパク質間の生物学的な関係性は相互作用ネットワーク・相互作用グラフとしてモデル化される。そのようなネットワーク中で、ノードは遺伝子やタンパク質に対応し、エッジはノードのペア間における相互作用を表現している。相互作用ネットワークは、密に接続された部分グラフを含む場合があり、密な部分グラフは「機能モジュール」として認識されている。 (Spirin and Mirny, 2003) 特に、Hartwell らは、生命現象の解析における機能モジュール間の相互作用の理解の重要性を主張している。

同じ部分グラフに属するタンパク質は、ある一つの特定の生命現象を引き起こすために利用されると考えられており、また、関連した機能を持つ傾向があると言われている。従って、巨大な相互作用ネットワークから機能モジュールの同定することは、遺伝子やタンパク質の機能を理解する上で基本的なステップの一つである。

相互作用ネットワークから機能モジュールを発見するというタスクはネットワーククラスタリング問題と呼ばれ、この目的のために様々なアルゴリズムが考案されてきた。例えば、Newman (2006) は、あるネットワークをクラスタリングし複数の部分ネットワークに分離した時、その分離がどれくらいうまくいっているかを示す指標 $Q$ を考案し、ネットワーククラスタリング問題を $Q$ の最適化問題として帰着させた。Newman らは彼らが開発したアルゴリズムを様々なソーシャルネットワークや生命ネットワークへ適用し、それらのコミュニティ構造の識別や比較を行った。生命科学分野においては、Bader と Hogue が 2003 年に MCODE と呼ばれる手法を開発した。MCODE はタンパク質間相互作用ネットワークからタンパク質複合体を発見するための手法である。MCODE は、初め、ネットワーク中のノードそれぞれに対し、ノードの接続性を反映したスコアを定義し、次に、そのノードのスコアが高いノードのグループを検索する。違う方法としては、Adamcsek らは clique-percolation (CPM)法を用いる手法を考案した。この手法では k−clique（k 個のノードを含む完全グラフ）を検索し、得られた k-clique を統合することで最終的なネットワーククラスタリング結果を得る。CFinder は、MCODE など他の手法に比べ、より密に接続された部分ネットワークを検出することが可能である。Rivera らは、NeMo と呼ばれる手法を提案した。これは、ネットワークに含まれるノードに対し、ノードの接続性を反映するログオッズスコアを算出し、このスコアを用いて、階層的クラスタリング法を行う手法である。Yanjun ら・Lei らはタンパク質複合体や機能モジュールを検出する目的で、機械学習的な手法を用いた。(Qi *et al.*, 2008; Shi *et al.*, 2011) 両者は、ネットワークのトポロジカルな特徴と既知のタンパク質複合体の生物学的な特性を組み合わせ、学習を行っている。

## 1. 大規模相互作用ネットワーク中に含まれる機能モジュール解析手法の開発

ネットワーククラスタリングはコンピュータサイエンス分野においてもっとも熱気のある分野の一つである。近年でもいくつかの新しい手法が開発・発表され続けている。(Andersen *et al.*, 2016; Seok-Joo Lee *et al.*, 2016; Zhang and Zou, 2015) これらの手法は、それぞれ異なった特徴を持っており、興味深い生物学的なモジュールを発見できる可能性がある。しかしながら、モジュール間のオーバーラップなど、幾つかの重要な相互作用ネットワークの特徴を取り扱うことができない手法が多い。

過去の相互作用ネットワークの調査報告 (Palla *et al.*, 2005; Rives and Galitski, 2003) によると、機能モジュールは同じノードを共有する場合がある、つまり、いくつかの機能モジュールはオーバーラップする部分を持つ場合があるということである。これはグラフ理論の分野の言葉で表現すると、相互作用ネットワークから機能モジュールを同定するには、ソフトクラスタリング手法が必要であるということである。

相互作用ネットワークにおけるソフトクラスタリング手法の重要性は、機能モジュールの core-peripheral 構造 (Gavin *et al.*, 2006; Kuhner *et al.*, 2009) の検討の重要性と表現することもできる。典型なシグナリングカスケードネットワークなど、2 つの強く関連したタンパク質群が多くのタンパク質を共有している場合、現在のハードクラスタリング手法を用いると、その 2 つのグループは一つの機能モジュールとして同定される。生物学の視点から考えると、core タンパク質はモジュールの中心的かつ重要な役割を果たし、また、peripheral タンパク質は、core タンパク質のうちのいくつかとのみ相互作用し、ネットワークの制御などのような、ある特定の機能に関与すると考えられている。既存手法では、このモジュール（クラスタ）間の関係性を効率的に明らかにすることができない。加えて、オーバーラップが許されていない手法では、複数の独立したクラスタが生成されてしまう。また、クラスタ間の関係性の可視化を可能にするネットワーククラスタリング手法も、ほとんど存在しない。表 1 に既存手法をまとめる。

表 1 既存ネットワークノードクラスタリング手法の特徴

| アルゴリズム | アルゴリズムに対する入力 | ソフトクラスタリングかどうか | Core-peripheralを検出するかどうか |
|---|---|---|---|
| Girvan-Newman | ネットワークトポロジー | ソフトクラスタリングではない | 検出できない |
| 階層的クラスタリング | ネットワークトポロジー | ソフトクラスタリングではない | 検出できない |
| MCODE (vertex weighting based method) | ネットワークトポロジー | ソフトクラスタリングである | 検出できない |
| CFinder (clique percolation based method) | ネットワークトポロジー | ソフトクラスタリングである | 検出できない |
| NeMo (odds ratio based method) | ネットワークトポロジー | ソフトクラスタリングである | 検出できない |
| MCL (flow based method) | ネットワークトポロジー | ソフトクラスタリングではない | 検出できない |
| ADMSC (spectral clustering) | ネットワークトポロジー | ソフトクラスタリングではない | 検出できない |
| ベイズ法を用いた手法 | ネットワークトポロジー、既知複合体の特性 | ソフトクラスタリングである | 検出できない |
| ニューラルネットワークを用いた手法 | ネットワークトポロジー、既知複合体の特性 | ソフトクラスタリングである | 検出できない |

## 1. 大規模相互作用ネットワーク中に含まれる機能モジュール解析手法の開発

既存手法に存在する制限を克服するため、NCMine と呼ばれる手法を開発した。この手法は相互作用ネットワーク中の機能モジュールの同定と、生物学的に意味のあるクラスタの抽出に焦点を当てたネットワーククラスタリング手法である。NCMine は (i) ノードの次数局所性を元にしたノードのスコアリングを用いることで「局所的な near-clique」を検出し、(ii) 局所的 near-clique をそれぞれのオーバーラップによって再起的に統合することで相互作用ネットワークから機能モジュールを発見する。(i) 局所クラスタの構築フェーズと (ii) その統合フェーズの 2 段階を用いることで、複数のモジュールに含まれるタンパク質群の検出と、モジュール間の関係性の検出両方を同時に可能にしている。

「near-clique」とは完全グラフからいくつかのエッジが欠けているグラフ構造を指す。完全グラフ（クリーク）に関することは数学的に深く研究されており、完全グラフの検出アルゴリズムなどはすでに広く知られており、比較的簡単である。しかしながら、near-clique の探索は完全グラフの探索と比較して大変困難である。

相互作用ネットワークからタンパク質複合体や機能モジュールを検出するために、完全グラフ構造検出アルゴリズムを用いた研究も複数存在する。 (Li *et al.*, 2005; Liu *et al.*, 2009) しかしながら、完全グラフは大変厳しい制約である。それは機能モジュールに含まれる要素のうち幾つかは、モジュールの限られた部分（モジュールに含まれるタンパク質のうちのいくつか）のみと相互作用する場合があることや、実験によるノイズの影響でいくつかの相互作用が消えたように見える場合があるからである。

Wu らは、初めて core-peripheral 構造の検索を行った。 (Wu *et al.*, 2009) その際、Wu らは、初めにクラスタの検出を行い、検出されたクラスタを拡張することで peripheral ノードの検索を行った。対して、NCMine ではクラスタの検出・統合を行い、クラスタのオーバーラップ部分を core として認識するという違いがある。詳しくは次節で述べる。

# 1.2 新規ネットワークノードクラスタリング手法の開発

## 1.2.1 NCMine アルゴリズム

NCMine アルゴリズムは 2 つの段階から構成される。まず第 1 段階では、ノード間の接続性から「複数の局所 near-clique クラスタ」が構築される。続く第 2 段階では、局所クラスタのオーバーラップを考慮し、局所クラスタがマージされる。

### 第 1 段階: 局所 near-clique クラスタの構築

クラスタリング対象（アルゴリズムの入力）として与えられたネットワークに含まれるノードそれぞれに対し、ノードの接続性を元にしたスコアが与えられる。このクラスタリング手法では、ノードの次数中心性がノードスコアとして用いられる。次に、最もスコアが高いノードを選択し（以後、これを初期ノードと呼ぶ。）、ノード一つが含まれる局所クラスタと考える。次に、局所クラスタと隣接するノードをスコア順に局所クラスタに加える操作を行う。局所クラスタと隣接するすべてのノードを走査するが、走査されたノードがクラスタに加えられるかどうかは、加える前の局所クラスタの cliqueness と加えた結果できあがる局所クラスタの cliqueness の差とあらかじめ設定された cliqueness-change の比較によって決まる。

cliqueness とはあるグラフ G と、G と同じノード数で構成される完全グラフの類似度であり次の式で定義される。

$$cliqueness = \begin{cases} (\#\text{ of existing edges in the graph}) / \\ \quad (\frac{1}{2}N(N-1))\ (N = \#\text{ of nodes in the graph}) \cdot \\ 1.0\ (\text{if } N \leq 1) \end{cases}$$

Cliqueness はクラスタとしてのまとまりの良さを評価するための重要な因子のうちの一つであるが、相互作用ネットワークに対して適用する際には、この cliqueness のみに依存してクラスタを同定すると問題が起こる場合がる。NCMine では cliqueness の他に、cliquenss-change と呼ばれる別な指標も局所クラスタの構築フェーズに導入している。Cliqueness-change はクラスタに”tail”が含まれてしまうことを避ける役割がある。”tail”ノードとは、クラスタに含まれるノードのうちのごく少数としかエッジを持たない（相互作用しない）ノードを指す。

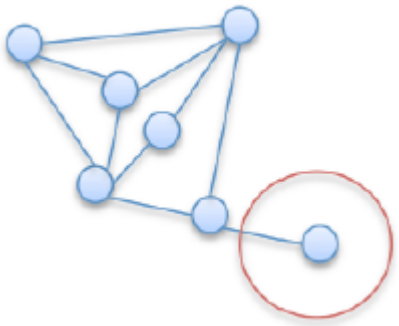


**図 2 tail ノードの例**

NCMine の局所クラスタ構築アルゴリズムでは、局所クラスタが大きくなると、いくつかの tail ノードがクラスタに取り込まれてしまうことがあるが、cliqueness-change の導入によりそれが防がれている。

局所クラスタはネットワーク中のすべてのノードをそれぞれ初期ノードとして構築される。また、この局所クラスタの構築は、ノードのスコアが高い順に行われる。構築された局所クラスタは続く第 2 フェーズで統合される。第 1 フェーズの計算量はおよそ $O(n^2)$（$n$はグラフに含まれるノード数）である。

### 第 2 段階: 局所クラスタの統合

第 1 段階で構築された局所クラスタは、局所クラスタの重なりを考慮して、再起的に統合（マージ）される。局所クラスタの重なりの度合いは、1 クラスタが 1 集合、クラスタに含まれるノードを集合に含まれる要素として考え、Jaccard 係数を用いて算出される。2 局所クラスタの重なりが閾値よりも大きく、かつ、統合後のクラスタの cliqueness が閾値よりも大きい場合、2 クラスタが統合される。しかしながら、2 局所クラスタの重なりが閾値よりも大きいが、統合後のクラスタの cliqueness が閾値に達しない場合、それらの 2 局所クラスタを階層関係があるクラスタ群として扱う。この場合、2 局所クラスタの共通部分が親クラスタ、2 局所クラスタの共通しない部分が、子クラスタとして階層関係が定義される。どの局所クラスタとも重なりが無いクラスタはそのまま最終的な解として出力される。

局所クラスタは局所クラスタ間の重なりの度合いによって統合され、また、1 ノードが複数のクラスタに属することがあり得る。したがって、本手法はソフトクラスタリング手法に分類される。

また、局所クラスタの統合は局所クラスタ生成順に行われる。それは、同じ near-clique に含まれるノードは、同じようなノード次数や次数中心性を持ち、また、類似した局所クラスタは局所クラスタ構築順が近いと期待されるからである。そのため、ノード次数中心性の性質と、局所クラスタ構築順より、局所クラスタの重なりの比較回数を削減できると考えられる。

この局所クラスタの統合フェーズを終えることで、最終的な解となるクラスタ群を得る。第 1 段階で構築される局所クラスタの最大数は$n$（$n$はグラフに含まれるノード数）であるため、第 2 フェーズの計算量はおよそ$O(n^2)$（$n$はグラフに含まれるノード数）である。

提案手法は Cytoscape (Smoot *et al.*, 2011) のプラグインとして実装を行った。この実装では、cliqueness-change とオーバーラップ・cliqueness の閾値の既定値はそれぞれ 0.2・0.6・0.6 と設定した。これは次節での人工ネットワークによるシミュレーションを元に設定した。

図 3 に NCMine アルゴリズムのフローチャートを示す。また補足図 45 に疑似コードを示す。

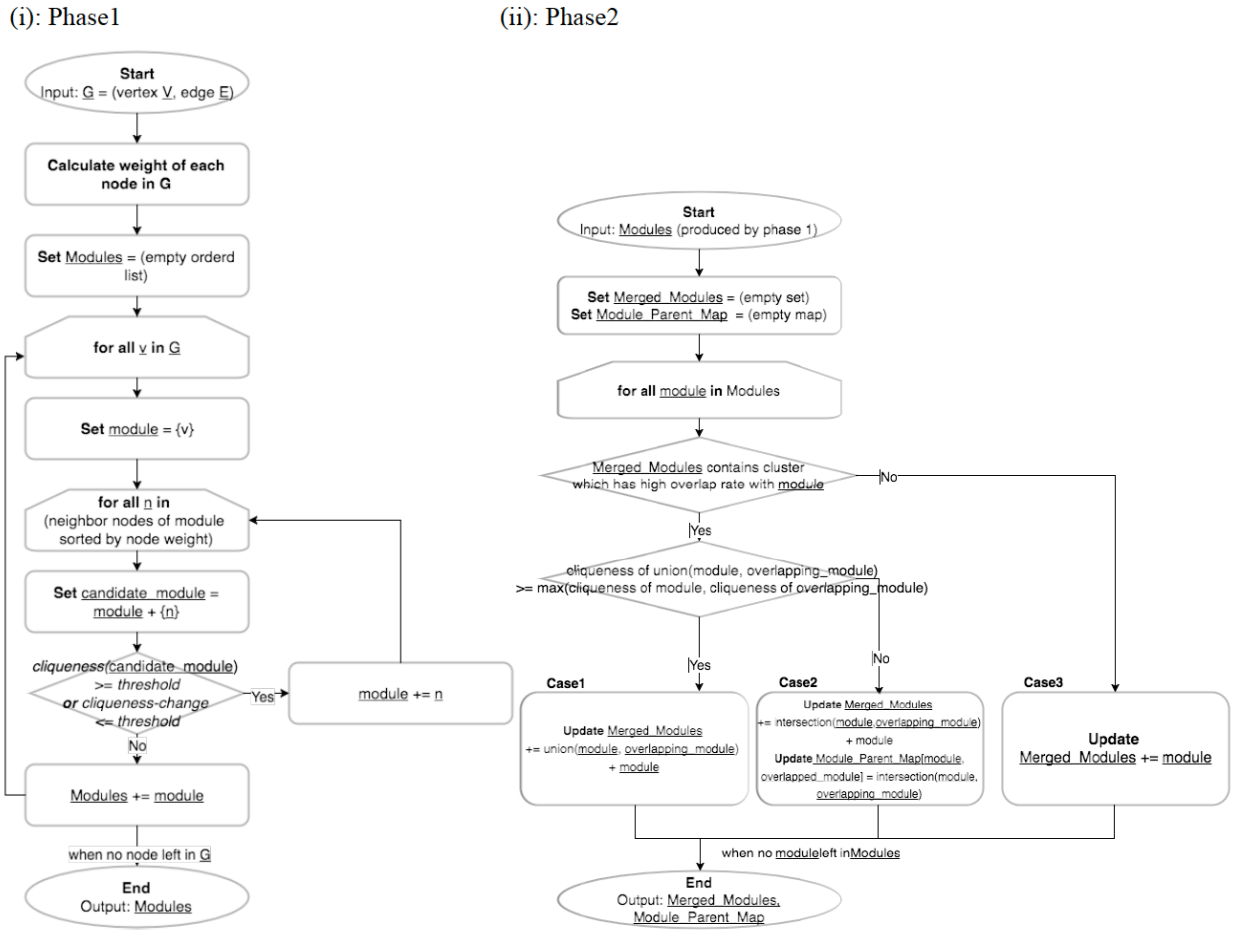



**図 3 NCMine アルゴリズムのフローチャート**

NCMine が最終的に返すクラスタリング結果は必ずしも大域最適解ではない。種々のネットワーククラスタリング手法はクラスタリング結果に関して考えることで、2 種類のカテゴリに分類することができる。それは、大域最適解を返す手法と、局所最適解を返す手法である。例えば、前節で触れた Newman の $Q$ は、ある 1 クラスタが他のクラスタからどれほどよく分離されているかを示す指標であり、実際のクラスタリングは $Q$ の最適化によって行われる。したがって、Newman 法は大域最適解を答えとして返す手法である。対して、NCMine はネットワーク全体からある限られた部分を抽出し、全体のサブセットから局所クラスタを構築する手法である。したがって提案手法は局所最適解を解として返す手法である。

## 1.2.2 人工ネットワークによるネットワーククラスタリング手法の評価

提案手法を評価するため、人工的にネットワークを生成し、生成したネットワークに対し、種々のネットワーククラスタリング手法を適用することで、各手法の評価・特徴の解析を試みた。初めに cliquness が 0.6 から 1.0 のランダムな near-clique を複数個生成し、次に、異なる near-clique の 2 ノードをランダムに選択し、その間にエッジを追加するという操作を複数回行うことで、near-clique が埋め込まれたネットワークを人工的に生成した。この人工的に生成したネットワークに対し、提案手法と既存の手法 (Bader and Hogue, 2003; Adamcsek *et al.*, 2006; Rivera *et al.*, 2010) をそれぞれ適用し、ネットワーククラスタリング結果を比較解析した。また、各ネットワーククラスタリング手法の出力と実際に埋め込まれた near-clique を比較することで、各手法の精度と再現率を算出した。2 クラスタ間の類似度は Jaccard 係数によって算出し、類似度が 0.6 以上のクラスタは同一のクラスタとして取り扱った。精度と再現率は次の式で算出される。

- 精度 = 真陽性率 / (真陽性 + 偽陽性)
- 再現率 = 真陽性 / (真陽性 + 偽陰性)

この式中で

- 真陽性: クラスタリングの出力に存在し、かつ、実際に埋め込まれた near-clique である
- 偽陽性: クラスタリングの出力に存在するが、実際に埋め込まれた near-clique ではない
- 偽陰性: クラスタリングの出力に存在しないが、実際に埋め込まれた near-clique 集合には存在する

## 1. 大規模相互作用ネットワーク中に含まれる機能モジュール解析手法の開発

提案手法の出力と既存の 3 手法の出力の比較解析を行った。既存手法の詳細に関してはそれぞれの論文にて述べられているが、ここでは、簡単に各アルゴリズムの概略を述べる。

MCODE (Bader and Hogue, 2003) アルゴリズムは (1) ノードの重み付け、 (2) ノード重みによるクラスタ構築、(3) 後処理の 3 段階から構成される。第 1 段階ではそれぞれのノードの重みが node-density と次数の積として算出される。続く第二段階では、ノードの重みが 1 番高い物が選択され、そのノードに周囲の重みが高いノードを追加していくことでクラスタが構築される。MCODE アルゴリズムでは、1 タンパク質が複数のクラスタに所属し得るが、同じタンパク質を共有するクラスタ間の関係性に関しては知ることができない。NCMine の局所クラスタ構築フェーズはは MCODE と類似した戦略を取っているが、局所クラスタを構成するための評価関数が異なり、また、MCODE の評価関数は、クラスタの中にハブタンパク質を取り込みがちで、そのため、やや大きいクラスタを生成する傾向にある。この特徴に関しては続く節で述べる。

CFinder (Adamcsek *et al.*, 2006) は CPM (Derényi *et al.*, 2005) アルゴリズムを用いて機能モジュールを検出する手法である。CPM アルゴリズムの基本は、ネットワーク中から極大クリークを検出し、それらを重なりによってまとめることである。CFinder は *k-clique* と呼ばれる *k−1* ノードを共有する部分グラフを解として出力する。この手法を遺伝子共発現やタンパク質間相互作用など、種々の相互作用ネットワークに対し適用したところ、clique がとても厳しい制約であると感じられた。それは、機能モジュールに含まれるいくつかのメンバータンパク質は、同じモジュールに含まれるいくつかのタンパク質のみと相互作用する場合が見受けられるためである。また、どの k-clique にも含まれないノードはどのクラスタにも含まれることはなく、また、グラフのノードやエッジによっては k-clique コミュニティを構成することができないという点が問題点としてあげられる。

NeMo (Rivera *et al.*, 2010) アルゴリズムでは $Rab$ と呼ばれるログオッズスコアを用いることで 2 ノード間（ノード *a* とノード *b*）のエッジの重みが算出され、これを用いて階層的クラスタリングを行うことで最終的な解を得る。複数の既存研究によると (Palla *et al.*, 2005; Rives and Galitski, 2003)、タンパク質複合体の形やそれを構成するタンパク質は状況に応じて変化することが示されているが、NeMo は 1 タンパク質を複数のクラスタに所属させることができず、そのため、生物学的相互作用ネットワークの特徴をネットワーククラスタリング結果に十分に反映させることができないと考えられる。

## 1.2.3 ヒトタンパク質間相互作用ネットワークを用いたネットワーククラスタリング手法の評価

人工ネットワークを用いた評価の他に、実際のヒトのタンパク質間相互作用ネットワークを用いて、種々のネットワーククラスタリング手法の評価・特性の解析を行った。ヒトタンパク質間相互作用ネットワークは Human Protein Reference Database (HPRD, Release 9) (Keshava Prasad *et al.*, 2009) より入手した。HPRD から入手したデータには、(i) 2 タンパク質間の相互作用データと、(ii) タンパク質複合体を構成するタンパク質のデータが含まれていた。この節の評価では(i) 2 タンパク質間の相互作用データに対し、種々のネットワーククラスタリング手法を適用し、得られた解を(ii)タンパク質複合体データを比較することで解析した。

ネットワーククラスタリング結果の比較解析に加え、解として得られたクラスタの生物学的な意味に関する検討も行った。得られたクラスタそれぞれに対し、Gene Ontology（GO）のエンリッチメント解析を行い、生物学的に意味のあるクラスタか、そうでないクラスタかの分類を行った。GO アノテーションは Gene Ontology Consortium (http://www.geneontology.org) より 2012 年 9 月に入手したものを利用した。GO タームのエンリッチメントは Fisher の正確確率検定を用いて算出し、p-value が 0.05 以下であるものを抽出した。また、多重検定の補正を行うため、Bonferroni 補正を行った。ネットワーククラスタリング手法の解として得られたクラスタのうち、一つ以上の enrich した GO タームが存在し、かつ、そのタームがクラスタを構成するタンパク質のうち 60%以上に付加されているクラスタを生物学的に意味のあるクラスタとして分類した。この比較は NCMine 以外にも、MCODE、CFinder、NeMo に対して行った。

### 利用データ・実験環境の詳細

遺伝子・タンパク質の機能アノテーションとして、Gene Ontology 以外に、The Kyoto Encyclopedia and Genes and Genomes（KEGG） (Kanehisa *et al.*, 2014) を用いた。

NCMine と比較対比する目的で、MCODE 1.4.0.beta2（2012 年 12 月リリース）、CFinder 2.0.5（2011 年 4 月リリース）、NeMo 1.4（2010 年 2 月リリース）を用いた。

すべてのネットワーククラスタリングのための演算は、同一のデスクトップコンピュータ (Apple iMac, 3.4-GHz Intel Core i7 CPU, 32-GB RAM)を用いて行った。

# 1.3 提案手法を用いたネットワーククラスタ解析

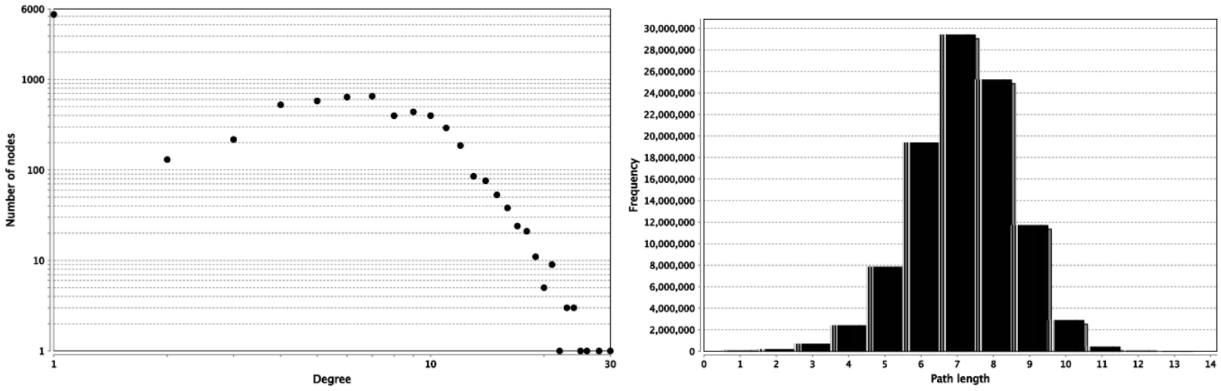
2.3.1 人工ネットワークを用いたクラスタリング手法の評価

構造的特徴がわかっているネットワークを人工的に生成し、各クラスタリング手法に対して適用することでクラスタリング手法のベンチマーク・特徴付けを試みた。

ここでは、10,000 ノードからなるランダムネットワークを生成し、さらにそれらのエッジをランダムに接続することにより、500 個のクラスタが埋め込まれたネットワークを生成する。この手順を用いて 1,000 個の独立した人工ネットワークを生成し、各クラスタリング手法に対して適用した。生成された 1,000 個の人工ネットワークの構造的特性は以下の通りである。

**表 2 人工ネットワークに含まれるクラスタ概要**

| | |
|---|---|
| **ネットワーク内クラスタのサイズ** | 7.20 ± 2.83 |
| **ネットワーク内クラスタの cliqueness** | 0.80 ± 0.11 |




**ノード次数分布** **２ノード間パス長分布**

**図 4 人工ネットワークのトポロジー特徴**

図 4 より、ノード次数がべき乗則に従い（スケールフリー性があり）、また、２ノード間のパス長も小さい（スモールワールド性がある）という性質が見受けられる。既存研究によると、実際のタンパク質間相互作用ネットワークも同様の性質を持っていることが知られている。(Barabási and Oltvai, 2004; Albert, 2005; Khanin and Wit, 2006)

NCMine を含む比較対象とするクラスタリング手法には、調整可能なパラメータを持つものがある。この実験ではそれらのパラメータを変えることにより、各クラスタリング手法のパラメータと出力結果の依存関係に関して考察した。各クラスタリング手法のパラメータ設定は表 3 の通りである。

**表 3 人工ネットワークに対し各クラスタリング手法を適用した際のパラメータ設定**

| | **規定値** | **ベンチマークに用いた値の範囲 (変化量)** |
|---|---|---|
| ***MCODE*** | | |
| Degree cutoff | 2 | 2 - 10 (2) |
| Node score cutoff | 0.2 | 0.0 - 1.0 (0.2) |
| K-core cutoff | 2 | 2 - 10 (2) |
| Degree cutoff | 2 | 2 - 10 (2) |
| ***NCMine*** | | |
| Cliqueness threshold | 0.6 | 0.0 - 1.0 (0.2) |
| Merge threshold | 0.6 | 0.0 - 1.0 (0.2) |
| Cliqueness-change threshold | 0.2 | 0.0 - 1.0 (0.2) |

表 3 で、例えば “Degree cutoff: 2-10(2)” は、Degree cutoff パラメータが 2 の場合, 4 の場合, …, 10 の場合を検討したということを示している。MCODE・NCMine は調整可能なパラメータ数が多いため、全ての可能なパラメータを試すことが大変困難である。従ってこの比較では、ある単一のパラメータを選択しそのパラメータの値を変化させるが、残りのパラメータはデフォルト値を使用するということを行った。

各クラスタリング手法に対し人工ネットワークを適用し得られた結果を表 4 に示す。

**表 4 各クラスタリング手法に対し人工ネットワークを適用した結果の概要**

| | 実行時間 (秒) | 抽出されたクラスタ数 | クラスタサイズ (平均±分散) | クラスタの cliqueness (平均±分散) |
|---|---|---|---|---|
| **NCMine** | 10 | 706 | 4–14 (6.77± 1.95) | 0.60–0.97 (0.67±0.06) |
| **MCODE** | 9 | 370 | 3–53 (7.72±5.84) | 0.09–1.00 (0.77±0.24) |
| **CFinder** | 10 | 1340 | 3–13 (8.41±2.52) | 0.57–1.00 (0.87±0.10) |
| **NeMo** | 90 | 953 | 4–13 (6.54±2.31) | 0.00–1.00 (0.57±0.37) |

表 4 より、ほぼ同等の実行時間（約 10 秒）でクラスタ抽出が完了していることが観察される。NeMo だけ突出して実行時間が長いように思われるが、これは NeMo のプログラミングとしての実装に問題があると考えられる。アルゴリズムとしての観点から考えると、計算量としては大きな違いは期待されない。

抽出クラスタ数を比較すると、CFinder が最も多くのクラスタを抽出し、対して、MCODE が最も少ないクラスタを抽出している。CFinder が用いている CPM アルゴリズムは、内部に “k-core”サイズと呼ばれるクラスタの大まかな大きさを決めるパラメータを持っており、CFinder は可能な k-core 全てを試しているため、その分、多くのクラスタの抽出を行っているように見えていると考えられる。MCODE はクラスタを同定する際に、主に、局所的ノード密度（ノード次数 * core-density）を用いているが、これは、放射状に接続されたノードをクラスタとして同定しやすい。したがって、MCODE は他手法と比較して、抽出されるクラスタのサイズは大きく数は少なくなると考えられる。表 4 よりその傾向が観察でき、また、ヒト PPI ネットワークを用いた実験でもこの傾向は一致している。

抽出されたクラスタのサイズ分布に着目すると、MCODE はサイズが 5~53（平均: 7.72、分散: 5.85）と様々なサイズのクラスタを抽出している。対して、NCMine, CFinder, NeMo は、ほぼ同じサイズ分布を示している。MCODE が抽出するクラスタのサイズ分布がブロードであるのは、MCODE が用いている評価関数が原因であると考えられる。MCODE はクラスタのサイズ分布だけではなく、クラスタの cliqueness 分布（0.09–1.00 (平均: 0.77、分散: 0.24)）も NCMine, CFinder と比較して広いという現象が観察される。この現象に関して MCODE の評価関数を用いて説明することはできないが、NeMo も同じ傾向を示している（0.00-1.00 (平均: 0.57、分散: 0.37)）。NCMine が出力するクラスタの cliqueness の分散は他手法と比較して、大変小さい。これは NCMine がネットワーク中から near-clique を抽出するのに特化しており、予め決められた cliqueness の閾値を満たす最も緩い局所的クラスタを抽出することを試みるからである。そのため、NCMine が抽出するクラスタのサイズは他手法と比較して小さく、また、クラスタの cliqueness の分散も小さいと予想される。これは表 4 に示される結果と一貫性がある。

次に、各クラスタリング手法によって抽出されたクラスタの概要だけではなく、クラスタそのものが実際に正解として埋め込まれたクラスタと一致しているかどうかを検証した。各クラスタリング手法のパラメータを変化させながらクラスタリングを実行し、各手法の出力として得られたクラスタと正解として埋め込まれたクラスタを比較し、検出率・適合率を算出した。

MCODE と NCMine はそれぞれ 4 個・3 個の調整可能パラメータが存在している。それぞれのパラメータを独立に変化させて試行を行うことは、パラメータ数が多く、大変困難であったため、ある一つの特定のパラメータの値のみを変化させ、残りのパラメータは表 3 に示されるデフォルト値に固定して試行を行い、検出率・適合率を算出することを行った。

この実験においても、前回の実験と同様に、1000 個の独立した人工ネットワークを生成し、検出率・適合率を算出した。1000 個の独立した人工ネットワークに対し、各クラスタリング手法のパラメータを変化させながら検出率・適合率曲線を描いたが、すべての人工ネットワーク

に対し、ほぼ同様の検出率・適合率曲線が描かれた。ここでは典型的な検出率・適合率曲線の一例を図 5 に示す。図 5 の 1 点がある一つのパラメータセットによって得られたクラスタ群から算出された検出率・適合率のペアを示している。 NeMo と CFinder は調整可能なパラメータが存在しないため、1 回の試行（1 点のみのプロット）となっている。

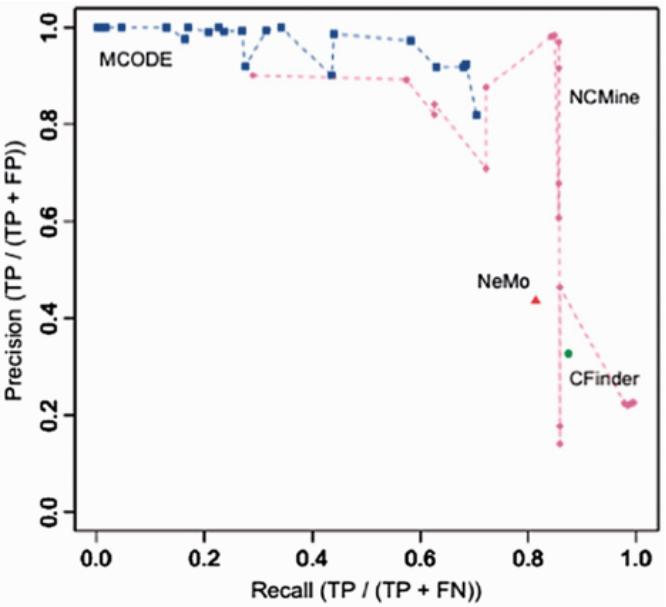



**図 5 人工ネットワークを用いた各ネットワーククラスタリング手法の評価**

図 5 より、NCMine は適切なパラメータを選択すると、near-clique 検出に関しては、検出率・適合率のよいトレードオフを示すことが観察される。cliqueness threshold = 0.6、merge threshold = 0.6、cliqueness change = 0.6 の場合に最大の適合率を得た。MCODE は高い適合率を出しているのに対し、検出力が低いことが観察される。高い適合率と低い検出力から、ほとんどの埋め込まれたクラスタが抽出されている一方、埋め込まれたクラスタのある一部のみしか発見されていないことが示唆される。それとは対照的に NeMo や CFinder は検出力が高く、適合率が低いことから、多くの偽陽性クラスタが生成されていることが考えられる。これは抽出クラスタ数（表 4）が多いこととも一貫している。

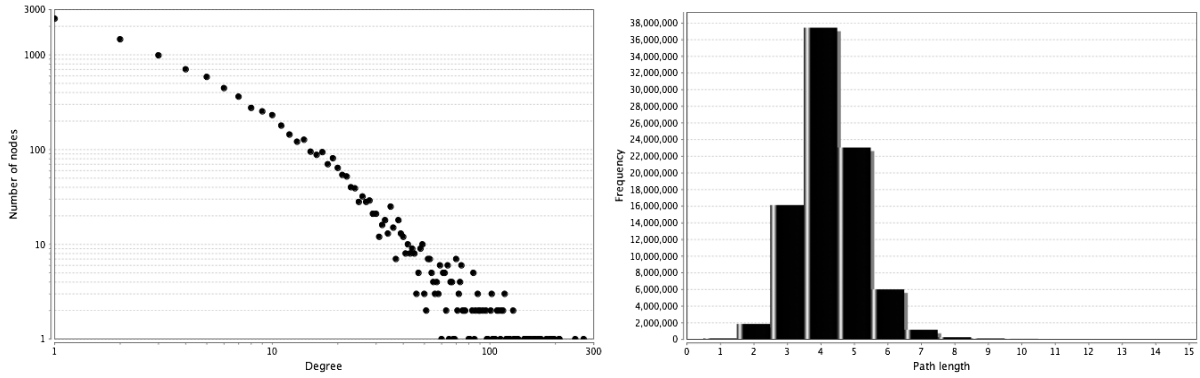
## 1.3.2 ヒト PPI ネットワークへの適用

次に、NCMine や既存手法を実際のヒト PPI ネットワークへ適用し抽出されるクラスタ群に関する検討を行った。抽出されるクラスタ群の評価は、抽出されたクラスタ群のうち「生物学的に意味のあるクラスタ」の割合を求めることで行った。ここで「生物学的に意味のあるクラスタ」は、クラスタに特異的に有意に出現する Gene Ontology term を持つタンパク質が含まれるかどうかで定義する。

PPI ネットワークは Human Protein Reference Database (HPRD) Release 9 (Keshava Prasad *et al.*, 2009) を利用した。このデータベースに含まれるタンパク質とその間の相互作用の数は表 3・図 6 の通りである。

**表 5 HPRD 9 に含まれるタンパク質数・相互作用数**

| | |
|---|---|
| **タンパク質数（ノード数）** | 9670 |
| **相互作用数（エッジ数）** | 39220 |




**ノード次数分布**　　**2 ノード間距離分布**

**図 6 タンパク質相互作用ネットワークのトポロジー概要**

## 1. 大規模相互作用ネットワーク中に含まれる機能モジュール解析手法の開発

HPRD Release 9 に対し、各クラスタリング手法を適用した結果を表 3 にまとめる。

**表 6 HPRD に対し各クラスタリング手法を適用して得られたクラスタの概要**

| | 実行時間 (秒) | 抽出クラスタ数 (A) | クラスタサイズ (平均±分散) | クラスタの cliquenes s(平均±分散) | 生物学的に意味のあるクラスタ数 | B / A |
|---|---|---|---|---|---|---|
| **NCMine** | 15 | 2309 (102*) | 3-21 (4.880± 2.887) | 0.600-1.00 (0.764±0.1 62) | 1259 | 0.54 |
| **MCODE** | 7 | 250 | 2-160 (9.359±19. 945) | 0.016- 1.000 (0.654± 0.337) | 93 | 0.45 |
| **CFInder** | 28 (10**) | 764 | 3-1010 (7.055±38. 954) | 0.017– 1.000 (0.900±0.1 64) | 315 | 0.41 |
| **NeMo** | 180 | 1510 | 4 - 68 (8.940±7.4 56) | 0.000- 1.000 (0.128± 0.244) | 784 | 0.52 |

* かっこ中の数値は core-peripheral 構造に含まれるクラスタの数を示す

** かっこ中の数値は CFinder の approx option を 0.1s に設定した場合の実行時間。

approx オプションによってある程度の探索の枝刈りが行われる。

表 6 より NeMo 以外は同等の実行時間で実行が完了していることが観察される。この観察事実は人工ネットワークによるクラスタリング手法の比較で得られた結果と一貫性がある。各クラスタリング手法の基本的な時間計算量はそれぞれ、NCMine・CFinder・NeMo が $O(n^2)$ 、MCODE が $O(nmh^3)$ である。したがって、実行時間の差は、各クラスタリング手法が行う前処理・後処理の差であったり、または、実装の違いによるものであると推察される。

MCODE は一番少ない数のクラスタを抽出しており、これは人工ネットワークでの結果とも一致する。しかしながら NCMine は実際のヒト PPI ネットワークに適用すると、他のクラスタリング手法よりも大変多くのクラスタを抽出している。NCMine はそのアルゴリズム上、パラメータとして与えられた閾値を満足するような near-clique を可能な限り抽出しようとする。したがって、実際の PPI ネットワークには大量のクリークのような部分グラフが含まれているということが考えられる。

クリークの様な部分グラフの重要性は、生物学的に意味のあるクラスタの割合によっても確認することができる。表 6 に記述されている生物学的に意味のあるクラスタの割合（B/A）は、NCMine が最大の値を記録している。CFinder は clqieness が比較的高いクラスタ（平均 0.9）を抽出しており、これは NCMine（平均 0.764）よりも高いが、生物学的に意味のあるクラスタの割合は、比較的低い（NCMine: 0.54、CFinder: 0.41）。この結果から、cliqueness が高いクラスタよりも、cliqueness がある程度のクラスタの方が生物学的な意味を考える上で重要であるということが考えられる。MCODE が出力するクラスタ群の中で生物学的に意味のあるクラスタの数は少なく、その割合も、他クラスタリング手法と比較した際に低いことが観察された。これは、MCODE がクラスタを抽出際に用いている評価関数が、複数のクラスタを一つの巨大なクラスタに結合しがちで、その結果、PPI ネットワーク中に存在する機能的関係性がないクラスタ群がまとめられてしまっているためであると考えられる。

次に各手法によって抽出されたクラスタ群と既知のタンパク質複合体との一貫性を検討した。この目的のために、HPRD よりタンパク質複合体データセットを入手し、1 タンパク質複合体が 1 クラスタであると考えた場合のクラスタの cliqueness 分布・サイズ分布を算出した。HPRD のタンパク質複合体データセットには、1521 個の既知のタンパク質複合体（複合体を形成するタンパク質のリスト）が記述されている。図 7 に各クラスタリング手法で得られたクラスタと既知タンパク質複合体の cliqueness 分布を示す。

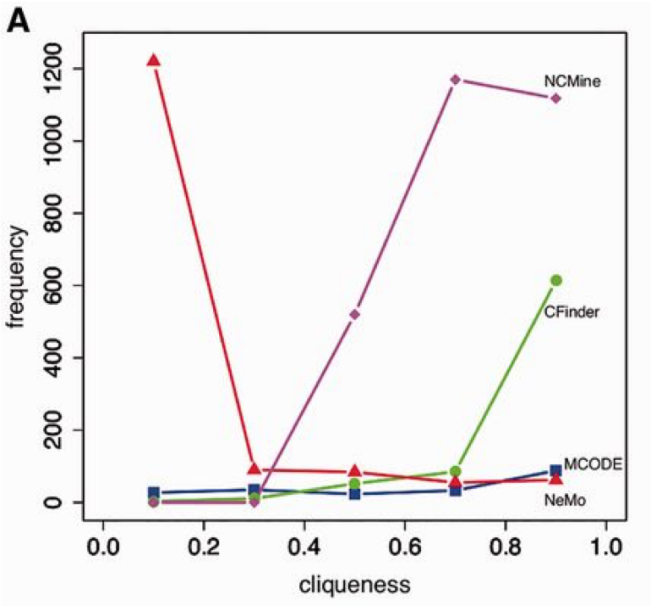



**図 7 HPRD データから抽出されたクラスタの cliqueness 分布**

図 7 を見ると、それぞれの cliqueness 分布が大きく異なっていることが確認される。既知のタンパク質複合体の cliqueness 分布は広く分散しているのに対し、各クラスタリング手法によって抽出されたクラスタ群から算出される cliqueness 分布は cliqueness が高い、または、低い場所に偏っている。対して、クラスタのサイズ分布は、クラスタサイズ = 3~5 に分布のピークがあるという点で、多くの手法が共通した特徴を持っている。また、図 8 に各手法により抽出されたクラスタのサイズの分布を示す。

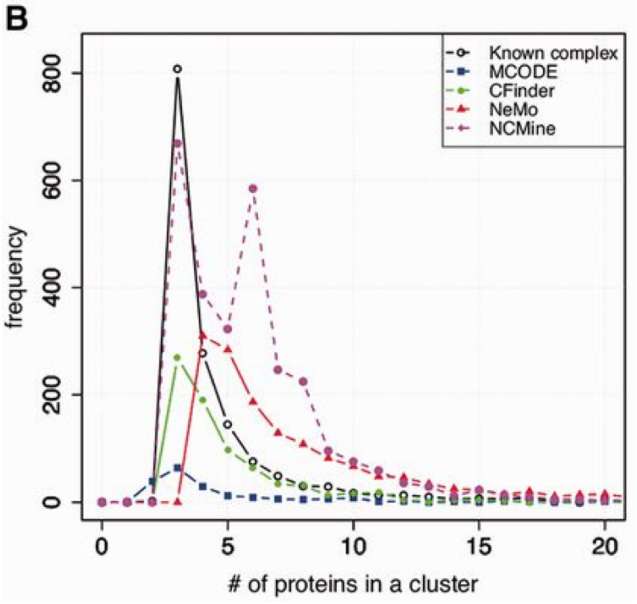



**図 8 HPRD データ抽出されたクラスタのサイズ分布**

NCMine はクラスタサイズ = 6 の部分に 2 個目のピークを持っているが、このピークを除くと、NCMine のクラスタサイズ分布と既知のタンパク質複合体分布のサイズ分布は大変類似していると考えられる。

分布の形状の違いの原因として、実験手法の違いや、実験データの取り扱いに関する違いが考えられる。タンパク質間の相互作用を検出するための実験手法と、タンパク質複合体を検出するための実験手法は、典型的に、異なっている。各クラスタリング手法の入力として与えられる PPI データは主に、タンパク質間の相互作用を調べる実験手法によって得られたデータであるのに対し、既知タンパク質複合体として扱っているデータはタンパク質複合体を検出するための実験手法によって得られたものである。また、タンパク質複合体を検出する為の実験手法で得られたタンパク質複合体を分解し、2 タンパク質間の相互作用として記述する場合もあり得る。その場合にも、タンパク質複合体を分解し 2 タンパク質間の相互作用として記述する方法も複数存在する。代表的なものに spoke model と matrix model が挙げられる。spoke model ではタンパク質複合体の中心的役割を果たすタンパク質が存在すると仮定し、そのタンパク質と複合体を形成する残りのタンパク質が相互作用するという分解が行われる。対して、matrix model では複合体を形成するすべての可能な 2 タンパク質間で相互作用が行われているかのように扱われる。このように実験手法やデータの取り扱いに関する違いによりクラスタの cliqueness 分布やクラスタのサイズ分布に違いが見えると考えられる。

## 1. 大規模相互作用ネットワーク中に含まれる機能モジュール解析手法の開発

NCMine の重要な特徴として、NCMine がソフトクラスタリングの手法であり、1 タンパク質を複数のクラスタに割り当てることが可能であることが挙げられる。図 9 にタンパク質のネットワーク中での次数と割り当てられたクラスタの数を示す。

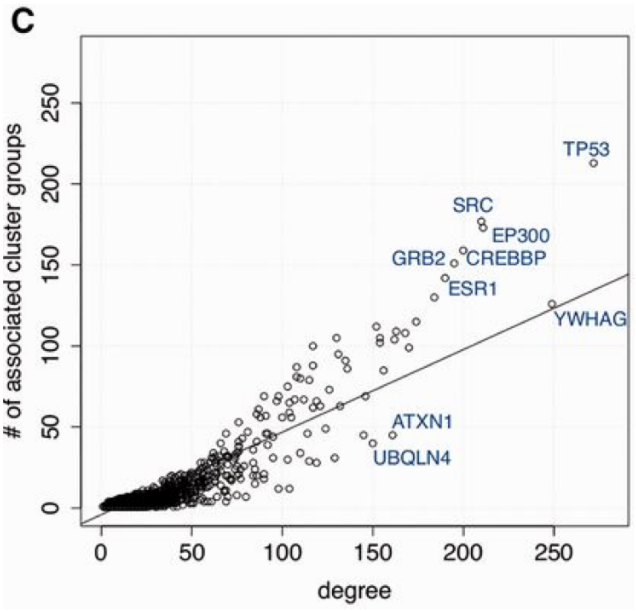



**図 9 ノード次数と所属クラスタ数の関係**

図 9 より、NCMine アルゴリズムは平均して、タンパク質が持つ相互作用数（ネットワーク中での次数）が増加するに従って、割り当てられるクラスタの数も増加していくことが観察される。次数とクラスタの割り当て数はほぼ線形であるが、いくつかの例外も観察される。図 9 中の直線は次数とクラスタの割り当て数から算出される回帰直線であるが、それよりも上に存在し、かつ、次数が高いものの多くはガンの進行に関わっていることがわかっているタンパク質であった（SRC、EP300、GRB2、CREBBP、RSR1）。ATXN1 や UBQLN4 のように回帰直線の下に存在するタンパク質は、割り当てクラスタ数が比較的少なく、また、ガンとの関連性も現在のところ確認されていない。この観察事実から、割り当てクラスタの数がタンパク質のガンとの関係性を示すよい指標になることが考えられる。

NCMine によってヒト PPI ネットワークから検出されたクラスタの例

タンパク質間相互作用ネットワーク中の core-peripheral 構造を見る利点を、実際に得られたクラスタを通して説明する。図 10 に得られたクラスタの 1 例を示す。

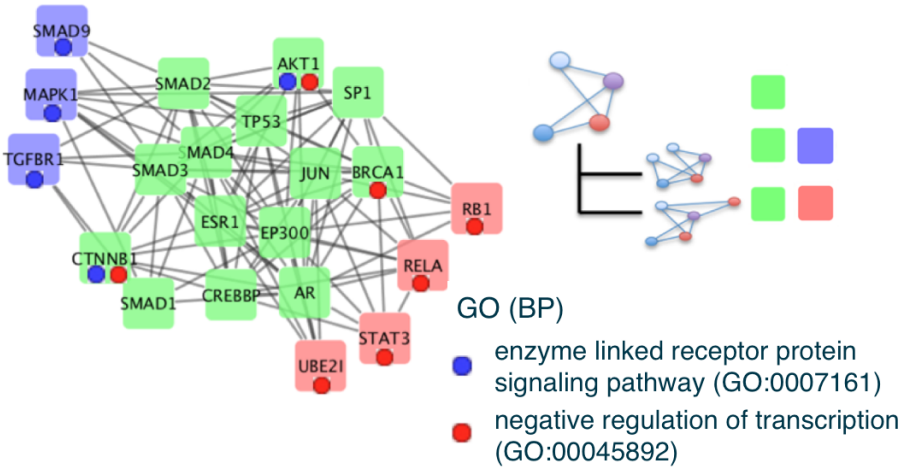



**図 10 NCMine によりヒト PPI ネットワーク中から得られたクラスタの例（1）**

**表 7 抽出クラスタ例 1 に含まれる遺伝子の一覧**

| 遺伝子名 | 説明 |
|---|---|
| AKT1 | AKT1 v-akt murine thymoma viral oncogene homolog 1 |
| AR | AR Androgen receptor |
| BRCA1 | BRCA1 Breast cancer 1, early onset |
| CREBBP | CREBBP CREB binding protein |
| CTNNB1 | CTNNB1 Catenin (cadherin-associated protein), beta 1, 88kDa |

| | |
|---|---|
| EP300 | EP300 E1A binding protein p300 |
| ESR1 | ESR1 Estrogen receptor 1 |
| JUN | JUN Jun proto-oncogene |
| MAPK1 | MAPK1 Mitogen-activated protein kinase 1 |
| SMAD1 | SMAD1 SMAD family member 1 |
| SMAD2 | SMAD2 SMAD family member 2 |
| SMAD3 | SMAD3 SMAD family member 3 |
| SMAD4 | SMAD4 SMAD family member 4 |
| SMAD9 | SMAD9 SMAD family member 9 |
| SP1 | SP1 Sp1 transcription factor |
| TGFBR1 | TGFBR1 Transforming growth factor, beta receptor 1 |
| TP53 | TP53 Tumor protein p53 |

図 10 は HRPD より得られたヒトのタンパク質間相互作用ネットワークに対し NCMine を適用して得られたクラスタのうち、ガンの進行に関係するダイナミクスが表されていると考えられる一例である。図 10 中の緑色のノードが親クラスタ（core クラスタ）にふくまれるタンパク質を表し、赤色のノードが子クラスタ（peripheral クラスタ）にふくまれるタンパク質を表す。図 10 の右上にクラスタの階層関係が示されている。また、KEGG パスウェイデータベースを参照することで、青色・水色・赤色のドットが打たれたタンパク質が、それぞれ、Wnt

signaling pathway、Pancreatic cancer pathway、Hepatitis B pathway 上に存在することを確認した。過去の知見によると、Wnt signaling pathway は胚発生に関係があり、また、様々なガンの発生・進行の鍵となるパスウェイであると考えられている。また、pancreatic cancer pathway や hepatitis B pathway はより特異的なガンの進行に関係するパスウェイである。Wnt signaling pathway に関係するタンパク質（青色のドット）は core タンパク質（緑色）にしか存在しないのに対し、Pancreatic cancer pathway（水色のドット）や Hepatitis B pathway（赤色のドット）に関係するタンパク質は peripheral タンパク質（赤色）にも存在することから、この図 10 に示されているクラスタの階層構造はガンの進行に関係があり、そのダイナミクスを示していると期待される。

HPRD より得られたヒトタンパク質間相互作用ネットワークから NCMine によって抽出されたクラスタの 2 例目を図 11 に示す。

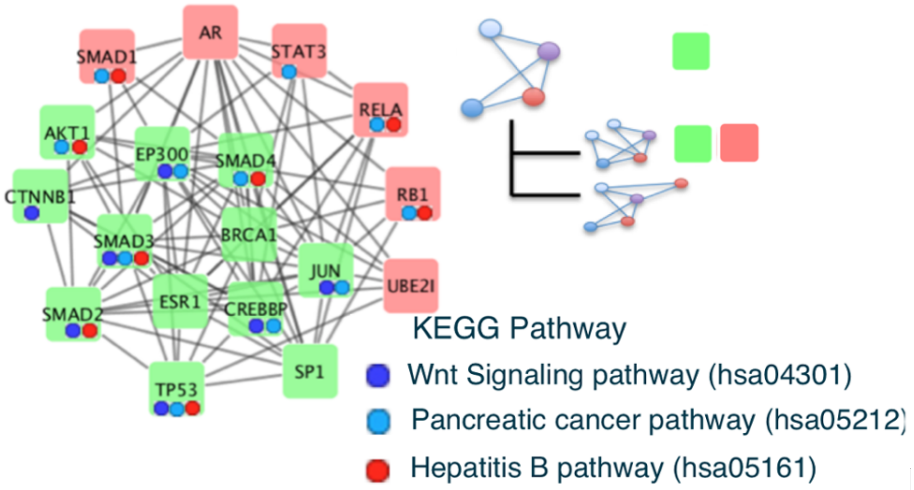



**図 11 NCMine によりヒト PPI ネットワーク中から得られたクラスタの例（2）**

表 8 抽出クラスタ例 2 に含まれる遺伝子の一覧

| 遺伝子名 | 説明 |
|---|---|
| AKT1 | v-akt murine thymoma viral oncogene homolog 1 |
| AR | Androgen receptor |
| BRCA1 | Breast cancer 1, early onset |
| CREBBP | CREB binding protein |
| CTNNB1 | Catenin (cadherin-associated protein), beta 1, 88kDa |
| EP300 | E1A binding protein p300 |
| ESR1 | Estrogen receptor 1 |
| JUN | Jun proto-oncogene |
| RB1 | Retinoblastoma 1 |
| RELA | v-rel avian reticuloendotheliosis viral oncogene homolog A |
| SMAD1 | SMAD family member 1 |
| SMAD2 | SMAD family member 2 |
| SMAD3 | SMAD family member 3 |
| SMAD4 | SMAD family member 4 |

| | |
|---|---|
| SP1 | Sp1 transcription factor |
| STAT3 | Signal transducer and activator of transcription 3 |
| (acute-phase | response factor) |
| TP53 | Tumor protein p53 |
| UBE2I | Ubiquitin-conjugating enzyme E2I |

図 11 において緑色で示されたノード（タンパク質）は NCMine によって発見された局所クラスタのうちの一つであり、また、緑色で示されたノードと赤色で示されたノードの集合も NCMine で発見された局所クラスタの一つである。これらの 2 局所クラスタが、局所クラスタの重なりによって統合され、緑色で示された部分が core、赤色で示された部分が peripheral として最終的な解として出力された。局所クラスタの親子関係は図 11 の右上に示される通りである。

一例目と同様に KEGG (Kanehisa *et al.*, 2014) パスウェイデータベースを検索すると、図 11 に示されるクラスタは膵臓ガン（pancreactic cancer pahtway; hsa05212; 青色のドット）と関係があることが示唆された。しかしながら、peripheral クラスタに見られる酵素結合レセプタータンパク質シグナリングパスウェイ（enzyme linked recepter protein signaling pathway; GO:0007167）と、転写の負制御（negative regulation of transcription; GO:0017148; 赤色のドット）が付加されたタンパク質は異なる分布を見せている。これらの結果から緑色で描かれるタンパク質は膵臓ガンの進行において中心的な役割を果たし、また、peripheral タンパク質は、それぞれ、異なるステージでその機能を果たしたり、また、別なパスウェイと関わりがあるのではないかと考えられる。

## ヒト PPI ネットワーク中に存在するタンパク質と機能モジュールの重なり

図 12 はヒト PPI ネットワーク中に存在する 1 タンパク質がどれくらいの機能モジュールに所属しているかを示している。各ネットワーククラスタリング手法をヒト PPI ネットワークへ適用し、それぞれのクラスタリング結果から計算されている。

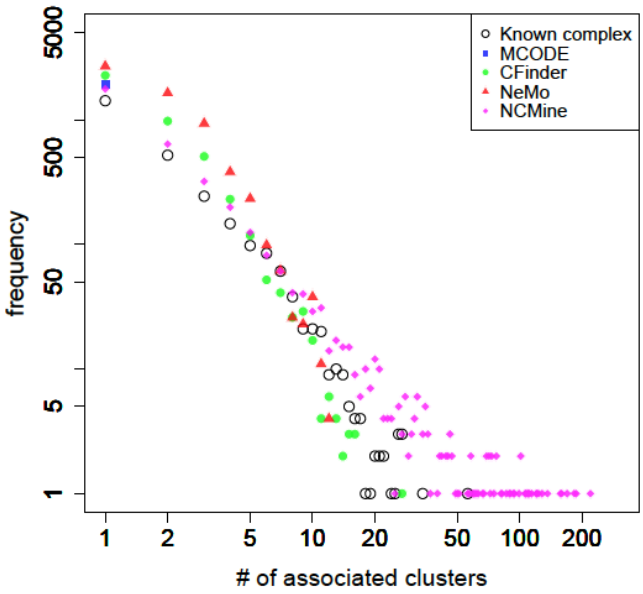



**図 12 ノードの所属モジュール数**

図 12 より、いくつかのタンパク質は、たくさんの機能モジュールに含まれていることが見て取れる。それらのタンパク質は相互作用する相手となるタンパク質が大量に存在するということが期待される。いくつかの既存研究によると、タンパク質間相互作用ネットワークには、ハブタンパクと呼ばれるタンパク質が数個存在するということが報告されている。2 種類のタンパク質が存在するということが報告されている。ネットワーク中のハブタンパク質の数はごく少数だが、相互作用パートナーの数は大変多い。つまり、ネットワークにおけるノード次数が大変高い。図はハブタンパクの存在を示していると期待される。

表 9 は NCMine によって検出された機能モジュールに含まれるタンパク質の中で、所属モジュール数が多いタンパク質を順に 3 個示したものである。表 10 は同様の値を既知タンパク質複合体データから算出したものである。

**表 9 NCMine によって検出された機能モジュールに含まれるタンパク質の中で所属モジュール数が多いもの**

| 所属モジュール数 | タンパク質名・説明 |
|---|---|
| 219 | TP53 tumor protein p53 |
| 187 | SRC v-src avian sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog |
| 56 | UBE2I ubiquitin-conjugating enzyme E2I |

**表 10 既知タンパク質複合体に含まれるタンパク質の中で所属複合体数が多いもの**

| 所属複合体数 | タンパク質名・説明 |
|---|---|
| 56 | ITGB1 integrin beta 1 |
| 34 | HDAC1 histone deacetylase 1 |
| 27 | ITGAV integrin alpha V |

p53 タンパク質は細胞周期を制御する転写因子であり、腫瘍抑制因子として働くことが知られている。また、SRC はガン遺伝子であり、遺伝子異常がガンを引き起こすことが知られている。ユビキチンはタンパク質のユビキチン修飾を通し、タンパク質の分解や転写制御など、様々な生命現象に関与するタンパク質であるが、UBE2I はユビキチンを修飾対象となるタンパク質へ結合させる時に重要な役割を果たす。integrin タンパク質は原形質膜に存在し細胞間の接着に関係するため、相互作用パートナーが多いのは当然のように思われる。

まとめると、機能モジュール抽出によって推定されるハブタンパク質は実際のハブタンパク質であり得る特性を持ち合わせていることが確認された。

# 1.4 NCMine の実装と公開

## Cytoscape プラグイン

Cytoscape（http://cytoscape.org/） (Smoot *et al.*, 2011) はオープンソースのネットワーク解析・可視化プラットフォームであり、バイオインフォマティクス分野において広く利用されている。Cytoscape コアは、ネットワーク（相互作用）データを扱うための基本的な機能、例えば、ネットワークデータが含まれたファイルを字句解析し読み込む機能・ネットワークに含まれるノードをレイアウトして表示する機能・ネットワークに含まれるノードやエッジを検索する機能などを提供している。Cytoscape コアは Cytoscape コア上に読み込まれたネットワークデータを操作するための API（Application Programming Interface）を外部に提供している。外部公開された API の利用により、Cytoscape の機能を拡張することが可能である。NCMine はこの API を利用することで、NCMine アルゴリズムを Cytoscape 上に読み込まれたネットワークに対して適用したり、ネットワーククラスタリング結果を可視化する機能を実装している。NCMine プラグインは Java 言語を用いて実装され、Cytoscape 公式のプラグインレポジトリである Cytoscape App Store（http://apps.cytoscape.org/） (Lotia *et al.*, 2013) にて公開されている。Cytoscape App Store に登録されているプラグインは Cytoscape のプラグインマネージャーよりインストールが可能である。図 13 は NCMine を用いたネットワーククラスタ解析の基本的なワークフローを表している。図では NCMine を用いて入力となるネットワークデータから near-clique を検出し、さらに、クラスタの階層関係の可視化を行うタスクを示している。NCMine の簡単なチュートリアルは Cytoscape App Store ページから確認することができる。

図 13 に NCMine Cytoscape プラグインを使ったネットワークノードクラスタリング解析のワークフロー例を示す。Cytoscape に読み込まれたネットワークに対し、NCMine アルゴリズムを適用し、抽出された機能モジュールを階層的に表示することができる。また、機能モジュール一覧に対し興味があるタンパク質で絞り込みを行うことや、抽出された機能モジュールを選択し、階層構造が分かりやすいように可視化することができる。

1. 大規模相互作用ネットワーク中に含まれる機能モジュール解析手法の開発

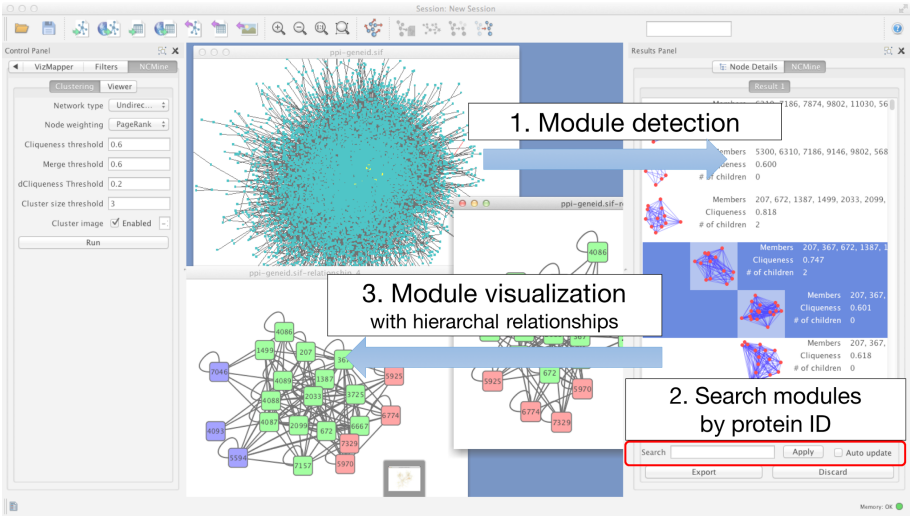



**図 13 NCMine Cytoscape プラグイン利用例**

## ATTED-II・COXPRESdb でのサービス提供

NCMine は、遺伝子共発現データベースである ATTED-II (Aoki *et al.*, 2016) ・COXPRESdb (Okamura *et al.*, 2015) にて利用可能な Web サービスとしての実装も行っている。図 14 に利用例を示す。

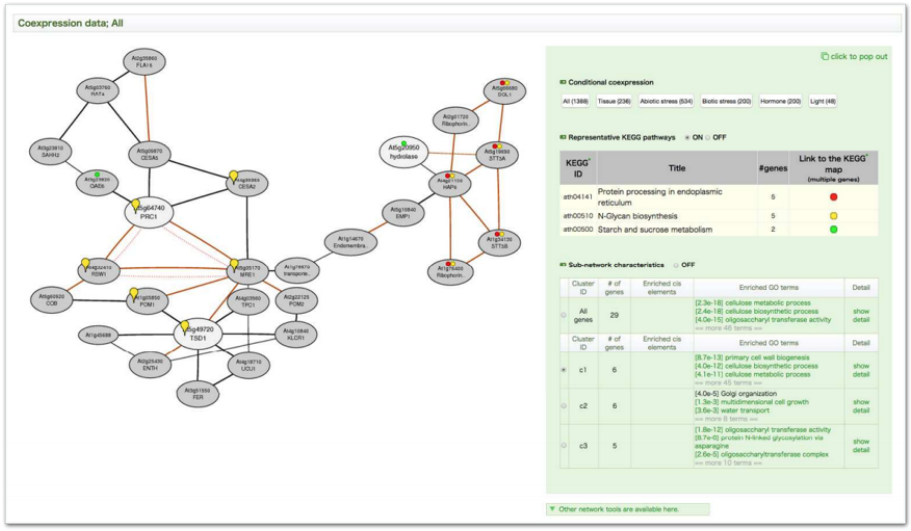



**図 14 ATTED-II での NCMine 利用例**

## 1. 大規模相互作用ネットワーク中に含まれる機能モジュール解析手法の開発

ATTED-II・COXPRESdb は遺伝子共発現のデータベースである。遺伝子共発現は、遺伝子ごとに、発現の組織特異性による発現プロファイルを作成し、任意の 2 遺伝子の発現プロファイルの類似性から遺伝子の関連度を算出する手法である。共発現度が高い遺伝子ペアを抽出することにより、遺伝子共発現ネットワークを描くことができる。

ATTED-II・COXPRESdb 上で NCMine は Web アプリとして利用可能である。入力として利用者が興味を持っている遺伝子を受け付け、その遺伝子と強く共発現する遺伝子を探索することで共発現ネットワークを作成する。作成された共発現ネットワークに対し、機能モジュール検索を実行することで、共発現ネットワークの機能理解を促進させるという効果を狙っている。この Web アプリケーションでは、機能モジュール抽出を行う他に、抽出された機能モジュールに対し、Gene Ontolog (The Gene Ontology Consortium, 2015) や KEGG Pathway (Kanehisa *et al.*, 2014) の濃縮検定を行うことで、さらに深い解析ができるように工夫されている。Web アプリとして実装することで情報科学分野以外の、計算機に詳しくない、例えば実験系の生物学者なども最新のアルゴリズムを大変簡便にに利用することができるという点で意義があると考えている。

## 1.5 小括

相互作用ネットワーク中の機能モジュール抽出を行う新規手法 NCMine を開発した。これは機能モジュールの、特に、core-peripheral 構造に焦点を当てたネットワークノードクラスタリング手法である。NCMine のアルゴリズムは（i）ノードの重み付けを行い完全グラフに近いネットワーク構造を抽出し、（ii）抽出されたサブグラフをサブグラフ同士の重なりによって統合することから成る。他手法との比較解析では、NCMine がより良い結果を出すことが示された。また、実際のヒトのタンパク質間相互作用ネットワーク解析において、機能モジュールの core-peripheral 構造とガンの発生の関係性が見出せる例を得ることができた。

# 2. タンパク質間相互作用・遺伝子共発現統合ネットワーク解析

## 2. タンパク質間相互作用・遺伝子共発現統合ネットワーク解析

## 2.1 導入

遺伝子やタンパク質間の相互作用の解析は、生命機能の理解にとって大変重要なステップの一つである。それは様々な生命現象は基本的に複数の遺伝子・タンパク質の相互作用の結果として現れるためである。生命現象を遺伝子やタンパク質の相互作用という点から検討する目的で、今日までに、様々な相互作用ネットワークが考案されてきた。その一つがタンパク質間相互作用ネットワークである。これは、グラフ中のノードがタンパク質を表し、相互作用する 2 タンパク質間にエッジが引かれるネットワークである。つまり、物理的なタンパク質間の相互作用を直接グラフとしてモデル化したものである。タンパク質間の物理的な相互作用は実験によって確認される。

タンパク質間相互作用ネットワークの他にも遺伝子・タンパク質間の相互作用を記述するネットワークは存在する。例えば、遺伝子共発現ネットワークである。遺伝子共発現とは、遺伝子発現パターンの類似性を用いることで 2 遺伝子の類似性を求めようとする手法である。図 15 に遺伝子共発現度計算の概略を示す。

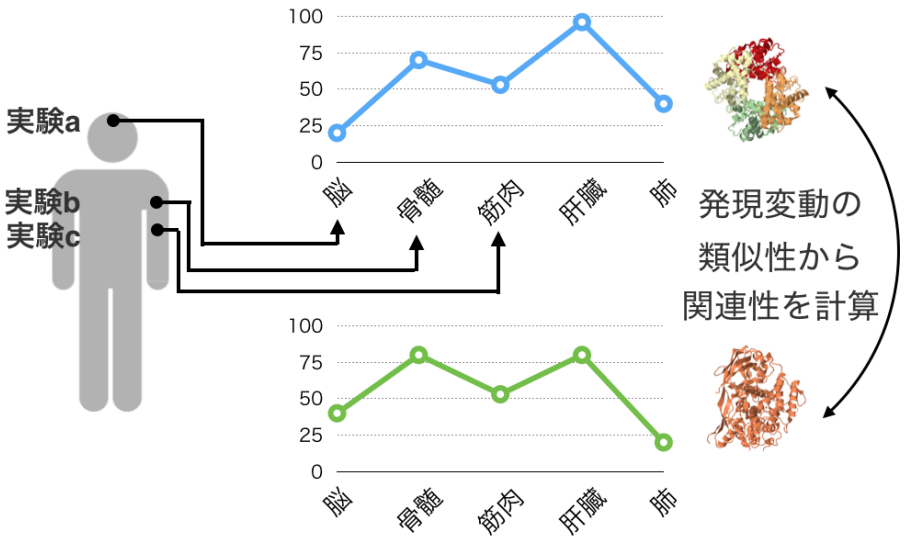



**図 15 遺伝子共発現度計算方法の概略**

遺伝子共発現は、遺伝子発現の組織特異性、つまり、どの組織で発現量が多く、どの組織で発現量が少ないかということを表す発現プロファイルを遺伝子ごとに作成し、そのプロファイルの類似性を相関係数を用いて算出することで 2 遺伝子の類似性を定量的に示す。遺伝子の発

## 2. タンパク質間相互作用・遺伝子共発現統合ネットワーク解析

現量の測定は、DNA マイクロアレイや次世代シーケンサーを用いた RNA-Seq と呼ばれる手法を用いて行われるが、このような実験では、一度に 1 組織（1 サンプル）の発現量しか計測することができない。1 遺伝子の発現プロファイル生成のためには、複数回の実験を行うことで複数組織（サンプル）で同じ遺伝子の発現量を計測し、それらのデータを統合する必要がある。DNA マイクロアレイを用いた発現量解析は、簡便さと定量性の面から広く利用されており、世界中の研究者によって測定されたデータは、NCBI GEO (Barrett *et al.*, 2013) や EBI ArrayExpress (Parkinson *et al.*, 2005) などの公共データベースに蓄積され、誰でも自由に利用することが可能になっている。公共データベースへのデータ登録件数も、年々増加の傾向を示しており、網羅的な遺伝子発現プロファイルの作成や、それを用いた、網羅的遺伝子共発現の算出を手軽に行うことができるようになっている。遺伝子共発現ネットワークは、遺伝子共発現の算出を行ったのち、遺伝子共発現度が高い遺伝子ペアを抽出し、ネットワークとして表現したものである。

タンパク質相互作用ネットワークと遺伝子共発現ネットワークはどちらも遺伝子やタンパク質の相互作用を元に生物を表現するネットワークであるが、様々な違いがある。図 16 に概略をまとめる。

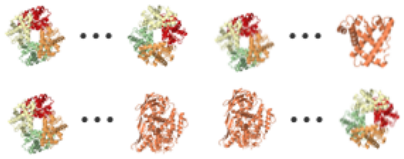
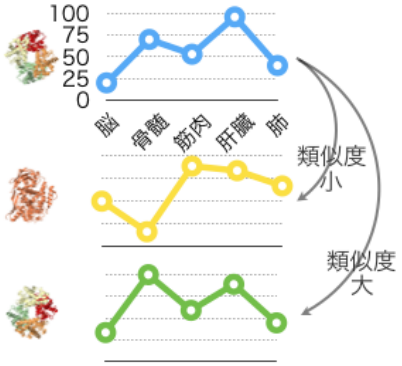
| | タンパク質間相互作用 | 共発現 |
|---|---|---|
| 概略 |  |  |
| 手法 | 実験的に2体間の相互作用を確認 | 発現パターンの類似性を計算 |
| 利点<br>・欠点 | ◎物理的な相互作用関係 | △非直接的な関係性 |
| | △全ての組み合わせを得ることは困難 | ◎網羅性が良い |

**図 16 タンパク質間相互作用・共発現の特徴**

## 2. タンパク質間相互作用・遺伝子共発現統合ネットワーク解析

例えば、タンパク質間相互作用は実験的に 2 タンパク質間の相互作用を確認するためその精度は高いが、遺伝子共発現は数多くの実験データを集約し発現パターンの類似性を算出するため、直接的な関係性を表すとは限らない。逆に考えると、タンパク質間相互作用を検出するためには 2 タンパク質ペアごとに相互作用があるかどうか確認する実験を行わなければならないため、時間と費用がかかるのに対し、遺伝子共発現は主に公共データベースに蓄えられているデータを用いて算出されるため、網羅性が良いという利点がある。

これらのことから、タンパク質間相互作用ネットワークと遺伝子共発現ネットワークなど、素性・利点などが異なるネットワークの重ね合わせを行うことで、それぞれの欠点などを補完できる可能性があると考えられる。このように複数種類の相互作用ネットワークを統合し、統合ネットワークより何らかの新しい知見を得ようとする既存研究はいくつか存在する。例えば String Database (Szklarczyk *et al.*, 2015) は、統合ネットワークをデータベース化し提供している。図 17 に String Database より提供される統合ネットワークの例を示す。

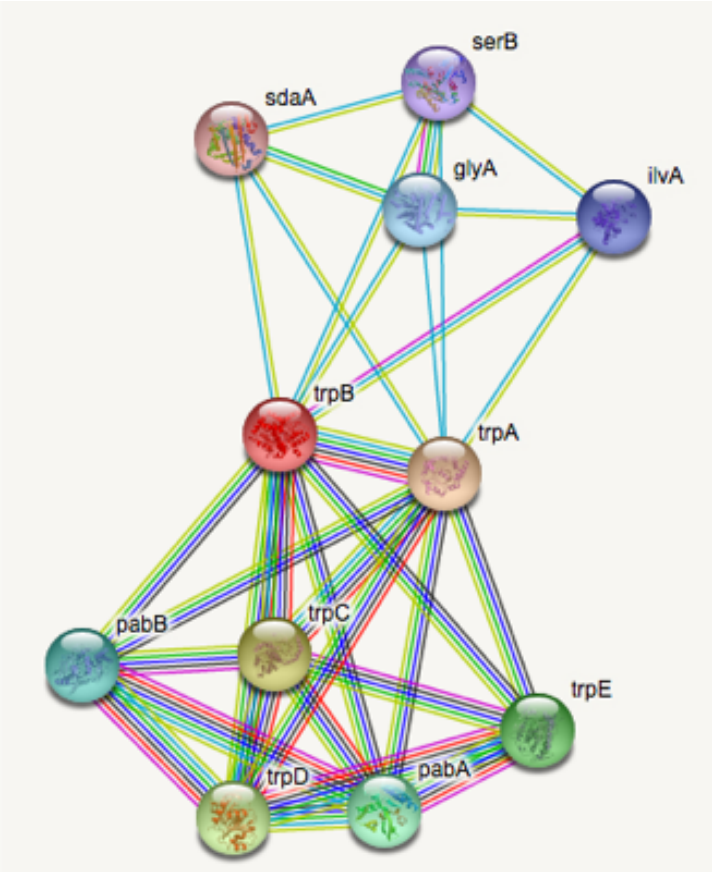



**図 17 String Database より提供される統合ネットワークの例**

## 2. タンパク質間相互作用・遺伝子共発現統合ネットワーク解析

図 17 では、円でタンパク質が描かれ、また、円と円の間の線がタンパク質間の相互作用を示している。円の間には複数の線が引かれ、線の色の違いが相互作用の種類の違いを示している。String Database では、統合ネットワークに描かれる相互作用（エッジ）として、タンパク質間相互作用や、遺伝子共発現、論文中における遺伝子名の共起関係など、複数種類の相互作用を提供している。

PTMapper (Narushima *et al.*, 2016) は、通常のタンパク質間相互作用ネットワークに対し、リン酸化情報を扶養するためのフレームワーク・アルゴリズムである。Narushima *et al.*, (2016) では、PTMapper を通常のタンパク質間相互作用ネットワークに適用し統合ネットワークを作成することで、通常のタンパク質間相互作用ネットワークだけでは発見することのできなかったパスウェイを見出すことに成功している。

このように、複数種類の相互作用ネットワークを統合解析することで、得られる知見を増やすといった試みが多々行われてきている。その一方、個々の相互作用ネットワークの質的・トポロジー的な違いや、その違いの生物学的な意義に関する検討が十分に行われているわけではない。第 2 章ではタンパク質間相互作用ネットワークと遺伝子共発現ネットワークの重ね合わせを行い統合ネットワークの作成を行う。その際、重ね合わせの元となるタンパク質間相互作用ネットワークと遺伝子共発現ネットワークのトポロジカルな性質に関して検討を行う。さらに、統合ネットワークに対し、解析の一例として、第 1 章で開発・検討を行った機能モジュール解析手法を適用し得られる機能モジュールに関する検討を行うことで、統合を行う効果などを分析する。

## 2.2 解析の概要と使用データ

### タンパク質間相互作用ネットワークと遺伝子相互作用ネットワークの特徴づけと統合

初めに、タンパク質間相互作用ネットワークと遺伝子共発現ネットワークを重ね合わせを行い、「統合ネットワーク」を作成する。タンパク質間相互作用データは Human Portein Reference Database（HPRD; Release 9） (Keshava Prasad *et al.*, 2009) を利用し、遺伝子共発現データは、COXPRESdb (Okamura *et al.*, 2015) より取得したヒトの共発現データ（Hsa3.c1-0）を利用する。COXPRESdb から入手できる遺伝子共発現データは、約 2 万個のヒトの遺伝子に対し、可能な全ペアの共発現度が含まれているため、共発現度に対し何らかの閾値を導入し、閾値を超える共発現遺伝子ペアを抽出することで遺伝子共発現ネットワークを構築する必要がある。そのため、閾値とする共発現度とその閾値によって得られる共発現ネットワークのサイズに関する検討も行う。タンパク質間相互作用と遺伝子共発現はタンパク質間・遺伝子間の相互作用を検出・計算する手法が大きく異なるため、タンパク質間相互作用ネットワークと遺伝子共発現ネットワークではノードの次数分布や 2 ノード間の距離分布など、トポロジー的な違いがある可能性がある。これらに関して検討する。最後に、タンパク質間相互作用ネットワークと遺伝子相互作用ネットワークの重ね合わせを行い、得られた統合ネットワークのトポロジーに関して比較検討を行う。

### 統合ネットワークに対する機能モジュールの抽出

構築した統合ネットワークより機能モジュールを抽出し、タンパク質間相互作用ネットワーク・遺伝子共発現ネットワークを統合する効果に関して検討する。機能モジュール抽出アルゴリズムとして前章で開発した NCMine を用いる。

# 2.3 タンパク質間相互作用ネットワークと遺伝子共発現ネットワークの比較

## 共発現ネットワークの作成・調整

Human Protein Reference Database（HPRD; Release 9）より取得したタンパク質間相互作用データは、9663 個のタンパク質、39220 本の相互作用を含んでいる。ここで利用する HPRD のデータは第 2 章で用いたものと同一のものであるが、遺伝子 ID の変換などの過程でいくつかの遺伝子が取り扱えなかったため、第 2 章で示した 9670 個から減っている。）対して、COXPRESdb より取得した遺伝子共発現データは全ての可能な遺伝子ペアの共発現度を含んでいる。そのため、共発現度に閾値を導入し、閾値を超える共発現度を持つ遺伝子ペアを抽出することで遺伝子共発現ネットワークを構築する必要がある。まず、閾値による遺伝子共発現ネットワークのサイズを確認した。COXPRESdb において共発現度の指標として Mutual Rank（MR） (Obayashi and Kinoshita, 2009) が用いられている。MR が小さいほど共発現度が高いということを示す。また、値の範囲は 1 から∞である。図 18 は MR による閾値を導入した際に生成される遺伝子共発現ネットワークのサイズ（ノード数・エッジ数）の変化を示している。

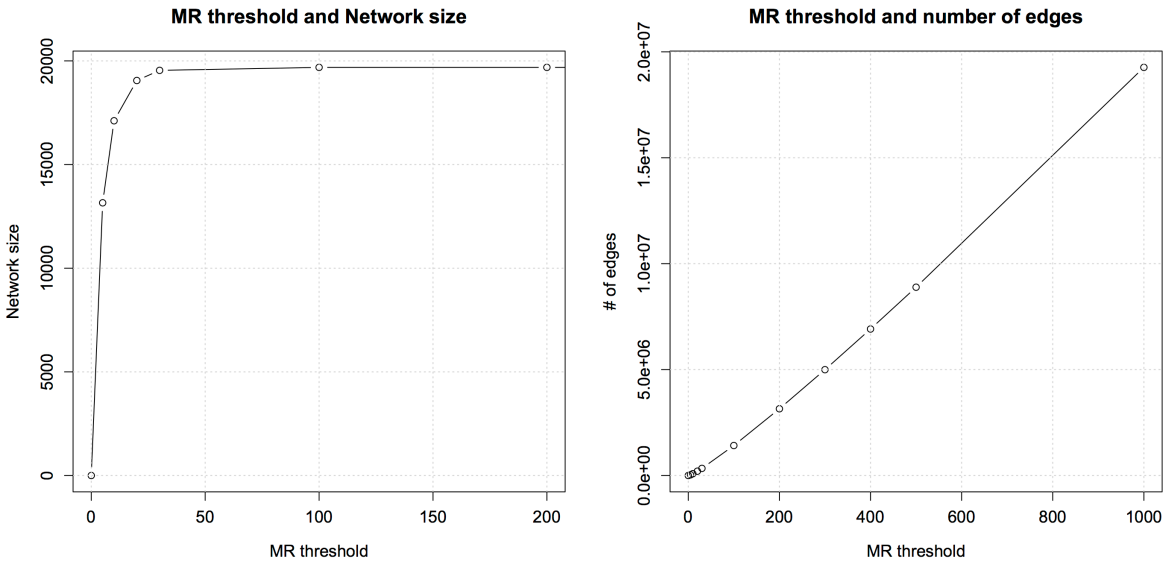



**図 18 共発現度による閾値と共発現ネットワークのノード数（左）とエッジ数（右）**

## 2. タンパク質間相互作用・遺伝子共発現統合ネットワーク解析

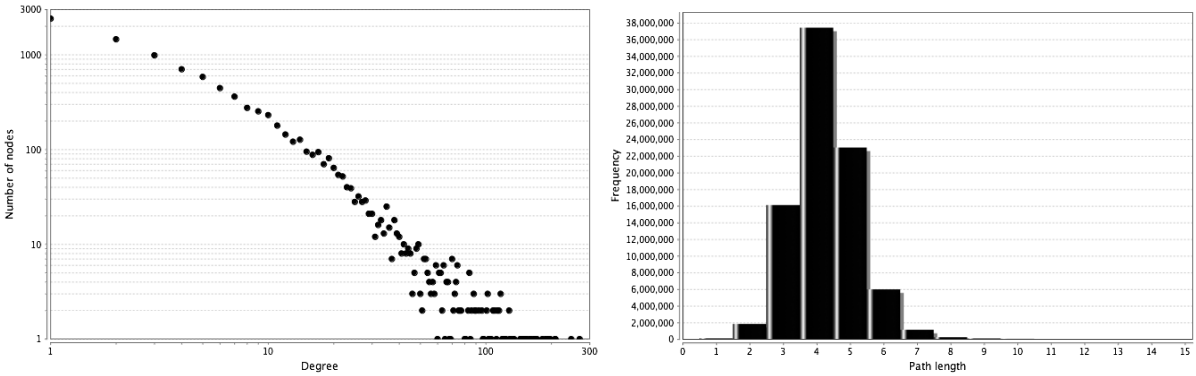
MR の閾値によるノード数の変化（図 18 左）より、MR の閾値を甘くしていくとあるところでノード数は飽和するような動きを見せている。ヒトの遺伝子数は約 2 万から 3 万であり、また、COXPRESdb に収録されているヒト遺伝子数は約 2 万であるため、MR を約 50 に設定することで生成される遺伝子共発現ネットワークにはヒトの遺伝子のほぼすべてが含まれることが示唆される。対してエッジ数は、MR の閾値と共に線形に増加している。（図 18 右）

この比較検証では、統合するタンパク質間相互作用ネットワークと遺伝子共発現ネットワークの影響の度合いを均等にするため、タンパク質相互作用ネットワークのノード数と同程度のノード数を持つ共発現ネットワークを統合の対象とすることにした。統合の元となるタンパク質間相互作用ネットワークと遺伝子共発現ネットワークの概要を表 11 に示す。また、それぞれのネットワークトポロジーの概要（ノード次数の分布・ノード間の最短距離の分布）を図 19、図 20 に示す。

**表 11 統合対象となるタンパク質間相互作用ネットワーク・遺伝子共発現ネットワークの概要**

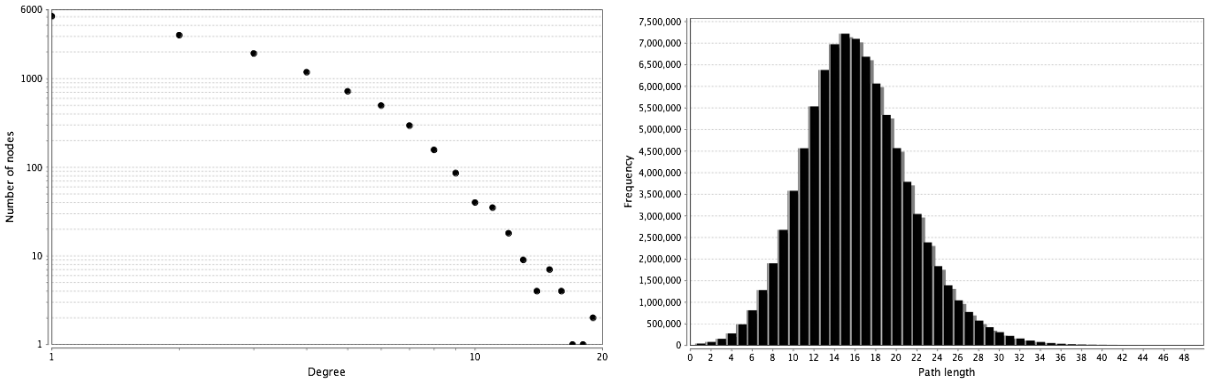
| | **ノード（タンパク質・遺伝子）数** | **エッジ（相互作用）数** |
|---|---|---|
| **タンパク質間相互作用ネットワーク** | 9663 | 39220 |
| **遺伝子共発現ネットワーク（MR <= 5)** | 13155 | 34302 |

## 2. タンパク質間相互作用・遺伝子共発現統合ネットワーク解析




ノード次数分布　　２ノード間距離分布

**図 19: タンパク質相互作用ネットワークのトポロジー概要**




ノード次数分布　　２ノード間距離分布

**図 20 遺伝子共発現ネットワークのトポロジー概要**

図 19、図 20 のノード次数分布を見ると、どちらもべき乗則に従っていることが観察される。従って両ネットワークは共通してスケールフリー性を持っていることが示唆される。対して、2 ノード間の距離分布は、分布の形は両者ともガウス分布を示しているのに対し、平均がそれぞれ、タンパク質間相互作用： 4、遺伝子共発現: 15 と大きく異なっている。遺伝子共発現ネットワークとタンパク質間相互作用ネットワークを比較した際、前者は後者よりも疎に接続されているということが示唆される。生物学的な視点から検討すると、タンパク質間相互作用は実験的に確認された直接的に相互作用するタンパク質を集めたネットワークであるのに対し、

## 2. タンパク質間相互作用・遺伝子共発現統合ネットワーク解析

遺伝子共発現ネットワークはタンパク質間相互作用よりも緩い関係性を示したネットワークであると考えられる。

図 21 にタンパク質間相互作用ネットワークと遺伝子共発現ネットワークを統合し他結果生成されたネットワークに含まれるノード・エッジの概要を示す。図 21 左が、統合ネットワークに含まれるノードがタンパク質間相互作用・遺伝子共発現どちらに含まれていたかを示すベン図である。図 21 右は統合ネットワークに含まれるエッジに対して同じものを計算した図である。

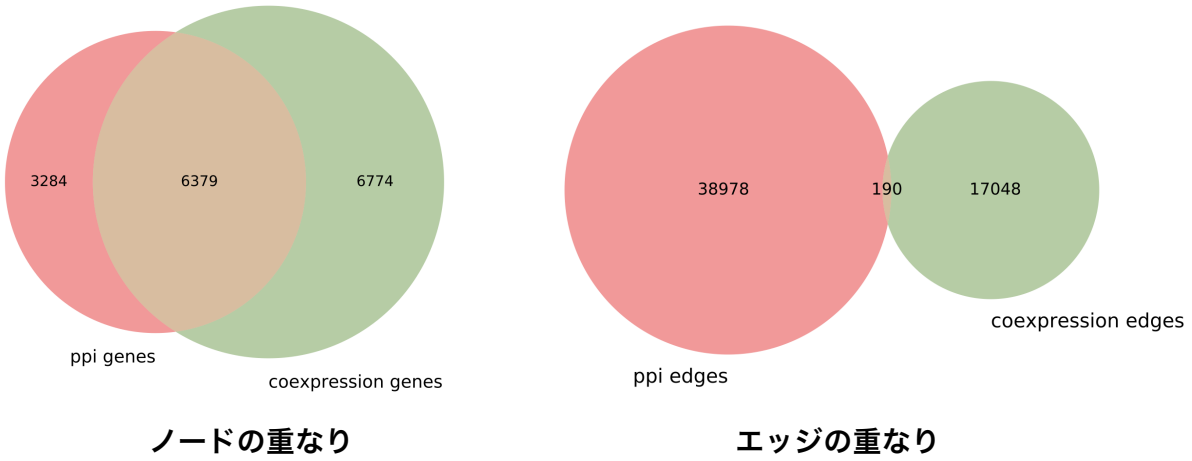
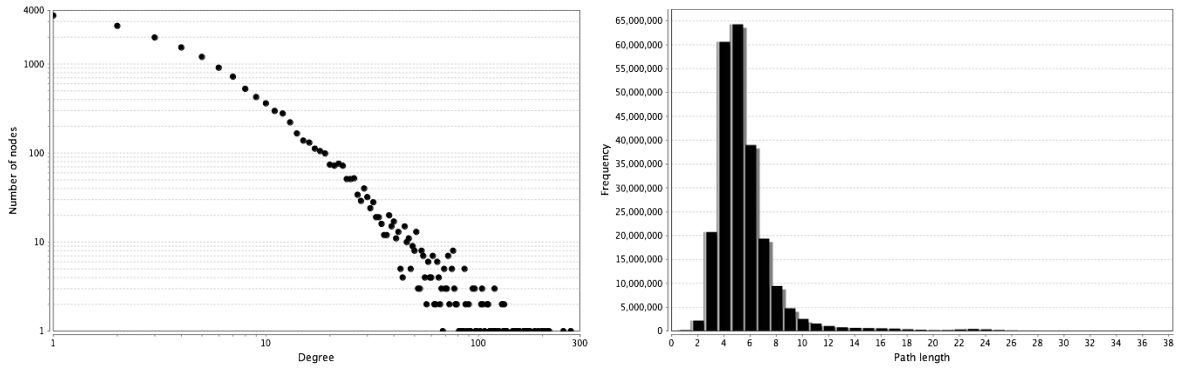



**図 21 統合ネットワークに含まれるノードとエッジの概要**

図 21 左は統合ネットワーク中におけるタンパク質間相互作用と遺伝子共発現のノードの重なりを示している。統合ネットワーク中には 3284 + 6379 + 6774 = 16437 ノードが存在し、そのうち約半分（タンパク質間相互作用 33%、の遺伝子共発現の 49%）が統合元の両ネットワークに存在する。対して、エッジ（タンパク質間の相互作用や遺伝子の共発現関係）に関しては、ごくわずかな重なりしか観察することができない。（図 21 右）これはそれぞれのネットワークを構築するための実験方法や計算方法に起因するものであると考えられ、また、同時にタンパク質間相互作用ネットワークと遺伝子共発現ネットワークはどちらも遺伝子・タンパク質間の相互作用を示すものであるが、わずかに異なった属性を表すネットワークであると考えられる。

図 22 に統合ネットワークのトポロジー概要を示す。




ノード次数分布　　2 ノード間距離分布

**図 22 統合ネットワークのトポロジー概要**

図 22 左のノード次数分布を見ると、統合後のネットワークのノード次数分布もべき乗則に従っており、スケールフリー性が確認される。また、統合ネットワークの 2 ノード間距離分布（図 22 右）では、分布のピークが長さ 5 の部分に見られる。統合前のタンパク質間相互作用ネットワークでは 4、遺伝子共発現ネットワークでは 15 であったことを考えると、共発現ネットワークにタンパク質間相互作用ネットワークを加えることで、単純な共発現ネットワークと比較して、ノードがより密に接続されるようになったということが示唆される。タンパク質間の相互作用関係と遺伝子共発現関係にあまり重なりが見られなかった点から考えても一貫した結果であると考えられる。

## 2.4 統合ネットワークに対する機能モジュール抽出

前節で構築した統合ネットワークに対し、NCMine アルゴリズムを適用し、機能モジュールを抽出行った。NCMine アルゴリズムのパラメータはデフォルト値（cliqueness-threshold = 0.6、merge-threshold = 0.6、cliqueness-change = 0.2）を用いて実験を行った。抽出されたクラスタは合計で 2743 個であり、そのうちタンパク質間相互作用・遺伝子共発現のエッジ両方が含まれるものが 1223 個（約 45%）であった。

ネットワーク統合の結果として興味深い例を挙げる。図 23 は統合ネットワークから抽出されたクラスタのうちの一つである。図のノードが遺伝子・タンパク質を示し、赤いエッジがタンパク質間相互作用、緑色のエッジが遺伝子共発現を示す。

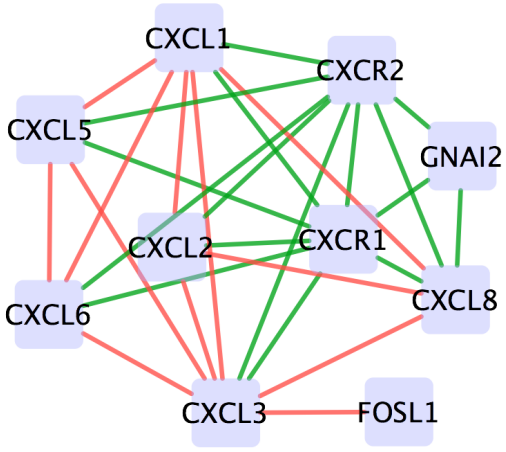



**図 23 統合ネットワークから抽出されたクラスタの例**

遺伝子名が CXCL から始まる遺伝子は CXC ケモカインに分類される。細胞の増殖・分化、細胞死に関係する情報伝達を行うタンパク質のことを総称してサイトカインと呼ぶが、その中でも白血球を遊走させる働きがあるものがケモカインと呼ばれる。CXC ケモカインはケモカインの中でもその配列中に CXC 呼ばれるモチーフを持つものである。CXC ケモカインは主として炎症の形成や血管新生などを行う働きがあることが知られている。 (Addison *et al.*, 2000) 遺伝子名が CXCR から始まる遺伝子は、CXC ケモカインのレセプターである。図 23 に示される統合ネットワークから抽出されたクラスタ例では CXCL 遺伝子群内は主にタンパク質間相互作用で接続されている。対して、CXCL 遺伝子群と CXCR 遺伝子群の間は遺伝子共発現で接続さ

れている。CXCR 遺伝子は CXCL 遺伝子のレセプターであるならば、二者は協調的に動作するということが強く考えられる。そのため、両者は一つのも機能モジュールに含まれることが望ましいと考えられる。

このクラスタの生物学的な意義を検討するため、KEGG Pathway (Kanehisa *et al.*, 2014) データベースの検索を行った。パスウェイとは代謝やシグナル伝達、遺伝子間の制御関係を経路図として示したものであり、KEGG パスウェイデータベースには、代謝やシグナル伝達などの様々なパスウェイ情報が登録されている。このデータベースより CXC サブファミリーを含むパスウェイを検索・取得した。下に取得したパスウェイ情報の一部を引用する。

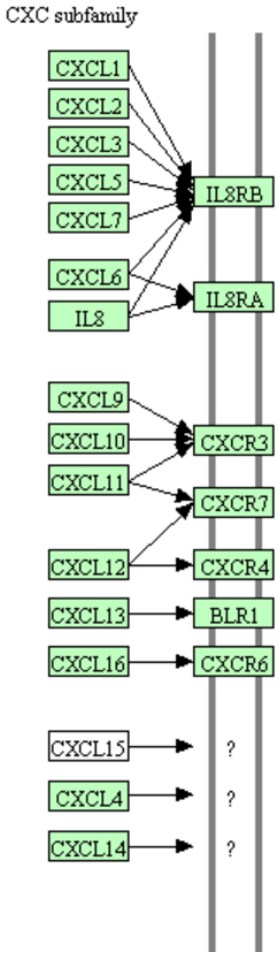



**図 24 Cytokine-Cytokine receptor interaction pathway (hsa04060）のうち CXC サブファミリに関係する部分**

## 2. タンパク質間相互作用・遺伝子共発現統合ネットワーク解析

図 24 中で緑色の四角がヒトに存在する遺伝子を示し、遺伝子間に引かれる矢印が遺伝子間にの相互作用を示す。IL8・IL8RB・IL8RA 遺伝子はそれぞれ CXCL8・CXCR2・CXCR1 遺伝子と同一の遺伝子である。図 23 と図 24 を比較すると、例として示した、統合ネットワークから抽出されたクラスタは CXC サブファミリが存在するパスウェイとよく類似していることがわかる。従って、タンパク質間相互作用ネットワークと遺伝子相互作用ネットワークを統合し、一つのネットワークとして解析することで、統合前のネットワークをそれぞれ個別に解析するよりも、より生物学的な知見を増やすことができる例を示した。

## 2.5 小括

本章では複数の種類の遺伝子相互作用ネットワークを統合した解析に関して検討を行った。本章では特に、タンパク質相互作用ネットワークと、遺伝子共発現ネットワークを取り上げた。タンパク質相互作用は直接的な相互作用を示すが網羅性に難点があるといった特徴を持っており、また、遺伝子共発現は網羅性では勝るが直接的な遺伝子関係を示しているとは限らないという特徴を持っており、両者は、利点・欠点が異なる。このように利点・欠点が異なるデータの統合により、お互いの補完になることを期待し検証を行った。

はじめに、タンパク質相互作用ネットワークと遺伝子共発現ネットワークのトポロジー的な差異に関して検討を行い、その生物学的な意味に関して考察を行った。次に、統合ネットワークから機能モジュール抽出を行った。その際抽出されたクラスタの中に、実際のパスウェイを再現していると思われるクラスタを見出すことができた。このクラスタには、タンパク質相互作用に対応するエッジと遺伝子興発現に対応するエッジ両方が含まれており、統合ネットワークからでなければ見出せないクラスタであった。

# 3. 大規模データ解析補助アプリケーションの開発

# 3.1 導入

近年、実験技術の向上やデータ解析用計算機の機能向上などのため、生命関連データの生産スピートが飛躍的に高まっている。例えば、図 25 はヒトゲノムを読む場合のコスト推移を表している。

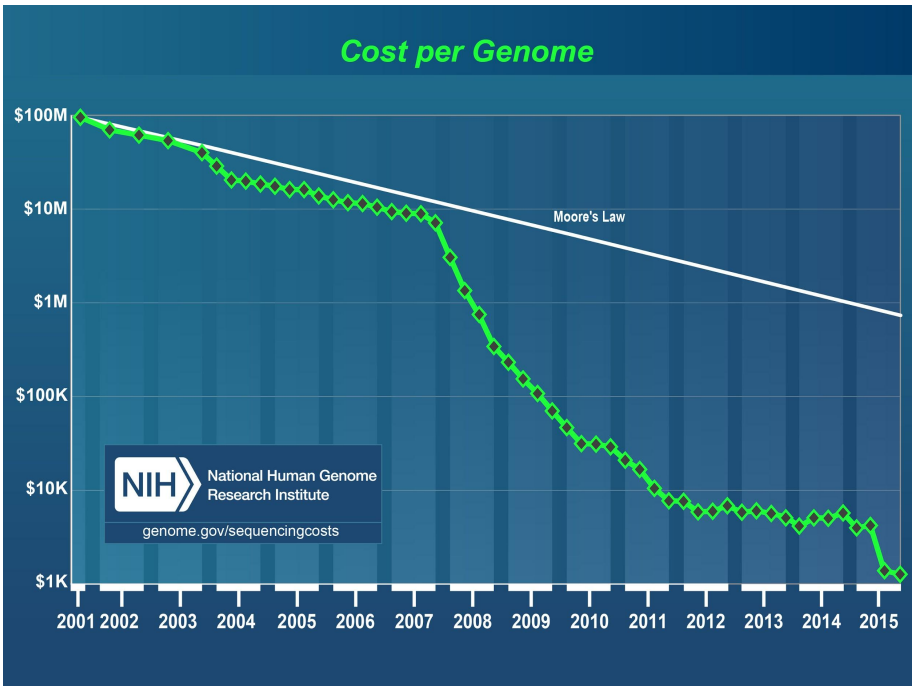



**図 25 ゲノム読み取りコストの年推移** (Wetterstrand, 2016)

2007 年終わりに米国 454 Life Sciences 社により次世代シーケンサーの先駆けとなる Genome Sequencer-FLX が発表され、ゲノムを読むコストが急激に下がった。また、2000 年ではヒトゲノムを読むためのコストが約 100 億ドルであったが、2015 年に入ると、約 1000 ドルにまで低下した。これは実験技術の向上のためである。

### 3. 大規模データ解析補助アプリケーションの開発

図 26 は NCBI GenBank に登録されている配列の数の年次変化を示している。

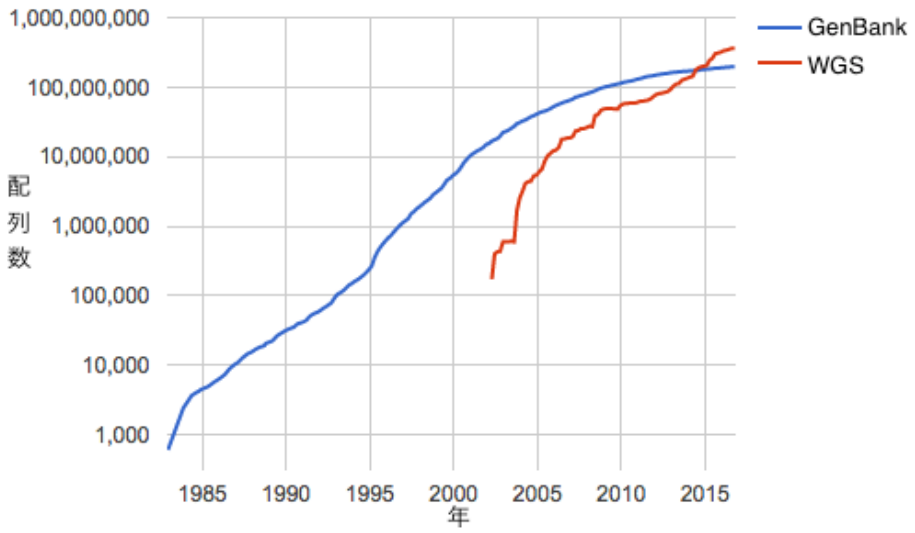



**図 26 NCBI GenBank に登録されている配列数の推移**
**（https://www.ncbi.nlm.nih.gov/genbank/statistics/ より作成）**

NCBI GenBank とは、アメリカ国立衛生研究所（National Institute of Health; NIH）によって運営されているデータベースであり、DNA や RNA などの核酸配列データを収録している。このデータベースに登録される配列データは、基本的に、世界中の研究者が自由に利用することができる。図 26 の GenBank とラベル付けられたラインが、NGS 以外の古いタイプのシーケンサー（Sanger シーケンサー）で解列が読まれたデータを示し、WGS とラベル付けられたラインが、NGS によって解読された配列データを示す。GenBank も WGS もどちらも年々増加の傾向を示している。また、近年、利用が進んでいる NGS データに関しては、既存のシーケンサーによるデータ増加のスピードよりも大変速いスピートでデータが増加していることが読み取れる。

GenBank は核酸配列のデータベースであるが、GenBank 以外にも大変多くの生命関連データを蓄えるデータベースが開発されており (Rigden *et al.*, 2016; Katayama *et al.*, 2013) 、また、それらに蓄えられているデータは増え続けている。増え続けるデータを人力で取り扱うことは大変な困難を伴う。コンピュータを用いて大量データを扱いやすい形で整理整頓し、実験生物学者などの利用者に提示できるようにすることは情報科学分野が生命科学分野にできることの一つであると考えられる。

「データを蓄える」ことは重要なことの 1 つであるが、それだけでは十分ではない。「蓄えられたデータからどのような意味を見出すか」ということも重要な事の一つである。しかしな

がら、データが増えるに従って、その増えたデータの取り扱いは次第に難しくなる傾向がある。例えば、近年の高速シーケンサーを用いると 1 回のシーケンサーの運転で約 500 ギガバイトものデータが算出される（Illumina HiSeq 200 の場合）。高速シーケンサーで得られた配列データの大規模解析を行おうとするなら数テラバイト・数ペタバイトものデータを扱う必要が発生する。また、データサイズの問題だけではなく、データの次元・自由度の問題もある。生命関連データでは特に、データを採取するために使った細胞の種類や状態などによって実験の結果が変わり得る。データから何かのパラメータを推定しようとする時、そのパラメータの数より、実験の結果として得られるデータが大変少ない場合がある（新 NP 問題）。この視点から情報科学分野の意義を考えた時、「計算機を使い、いかに蓄えられたデータを効率的に活用できるか」ということが重要な点のうちの一つである。

これらのことをまとめると、大量の増え続けるデータを管理・運用し、その結果として意味のあることを発見するための基盤プラットフォーム・アルゴリズムの開発が、情報科学分野として重要であると考えられる。

## 3.2 遺伝子ネットワーク解析アルゴリズム - NCMine

タンパク質の相互作用を解析することは、細胞機能の理解を行う上で大変重要である。それは、生命科学的な現象は、基本的に、単一のタンパク質の働きによって引き起こされる訳ではなく、複数のタンパク質の相互作用の結果として引き起こされるからである。例えば、図 27 は転写因子複合体を示している。

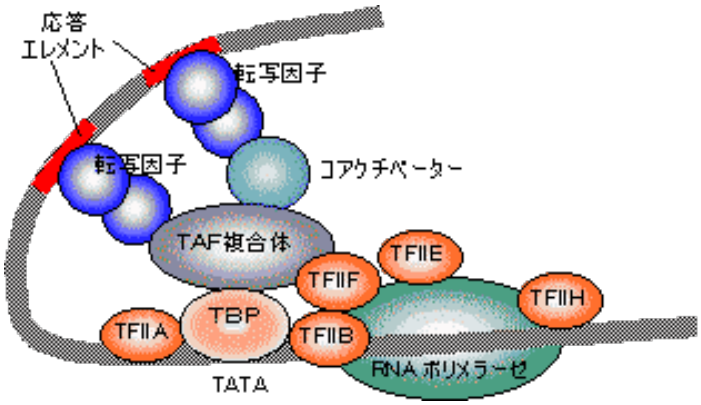



**図 27 転写因子複合体（http://www.sc.fukuoka-u.ac.jp/~bc1/Biochem/transcrp.htm より引用）**

図 27 に示されるように転写因子複合体は、TAF 複合体・TBP・TFII A・TFII B など様々な因子が関係している。また、これらの因子が同一空間に秩序なく存在しているだけでは転写因子複合体として機能することはなく、因子間の結合や相互作用が決まった状態を取らなければならない。従って、生体内で起こる分子機能の意味を検討する上で、1 分子の働きを検討することは重要であるが、加えて、複数分子間の相互作用も同時に検討することが重要である。しかし、遺伝子相互作用は複雑に関係し合っている。また、例えばヒトの遺伝子は約 3 万種類存在すると考えられており、それらの間のすべての相互作用を考えるならば、幅 3 万*高さ 3 万の大きさをもつ行列を考える必要がある。

遺伝子相互作用をモデル化する手法の一つとして、ネットワーク・グラフによる遺伝子相互作用の表現方法がある。グラフによる表現では、グラフ中のノードが遺伝子に対応し、ノード間にひかれるエッジが遺伝子間の相互作用を表現する。このようなグラフ表現を行うと、例えば、転写因子複合体など、ある特定の生体機能に関わる遺伝子に対応するノード間にはより多くのエッジがひかれ、対して、関係の薄い遺伝子間にはあまりエッジがひかれないことが期待

される。逆に考えると、遺伝子相互作用ネットワーク中から特に密に接続されている部分を抽出すると、それが、ある特定の生体機能を担う遺伝子群であることが期待される。

NCMine は遺伝子相互ネットワーク中からクラスタ構造を抽出することにより、ある特定の生体機能を担う遺伝子群、すなわち、機能モジュールを発見する手法である。NCMine は完全グラフに近いグラフ構造をクラスタと認識し、抽出を行う。アルゴリズムの詳細は、第 1 節 大規模相互ネットワーク中に含まれる機能モジュール解析手法の開発に記す。

NCMine は生命情報科学分野で広く用いられるネットワーク解析・可視化ソフトウェアである Cytoscape (Smoot *et al.*, 2011) のプラグインとして実装されている。NCMine の実装は Cytoscape プラグインのレポジトリである Cytoscape App Store (Lotia *et al.*, 2013) から入手可能である。NCMine は 2016 年の 1 月から公開を行い、2016 年 10 月末までで約 600 回ダウンロードが行われた。 図 28 このダウンロードは日本からのダウンロードにとどまらず、世界各国から行われている。

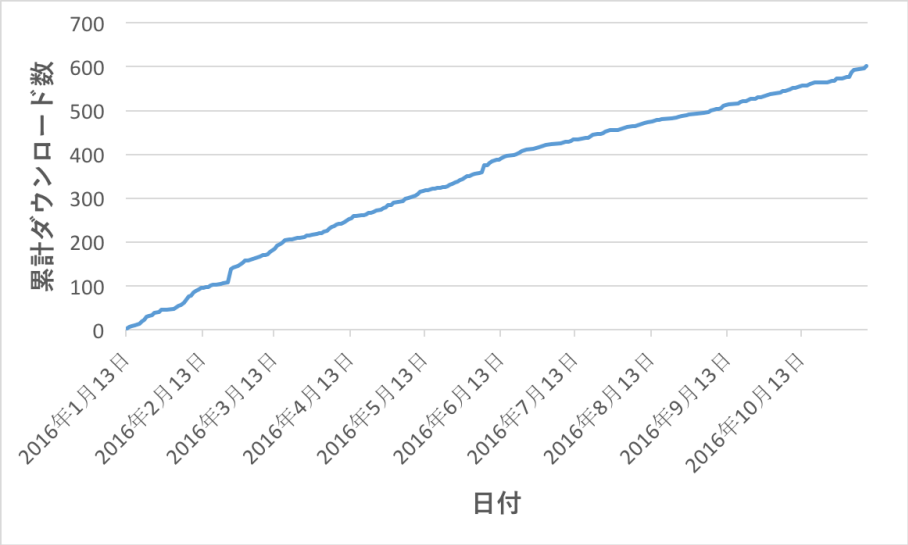



**図 28 NCMine プラグインのダウンロード数の推移**

## 3.3 遺伝子関係性データベース作成のためのフレームワーク PairDB の開発

遺伝子やタンパク質間の相互作用の解析、つまり、相互作用ネットワークの解析は生命現象のよりよい理解のための第一歩である。それは、基本的にあらゆる生命現象は、単一の遺伝子やタンパク質によってもたらされるわけではなく、複合体や機能モジュールのような複数の遺伝子が協調して動作することによりもたらされるからである。

遺伝子共発現やタンパク質間相互作用などの遺伝子ペアの関係性データの量は年々増加しており、ATTED-II（http://atted.jp） (Aoki *et al.*, 2016) や BioGRID（https://thebiogrid.org/） (Oughtred *et al.*, 2016) のような相互作用データを蓄積するデータベースが数多く構築・運用されている。しかしながら、相互作用データベースを構築することは簡単ではない。それは、相互作用データベースを構築するためには、全ての遺伝子ペア間の相互作用強度を保持する巨大、かつ、密な行列を効率的に管理することが求められるからである。また、遺伝子ペアではなく、単一遺伝子のデータベースを構築する際は、MediaWiki（https://www.mediawiki.org）など、様々なフレームワークを使い簡便に構築を行うことが可能であるが、遺伝子ペアのようなデータを格納するためのフレームワークはほとんど存在していない。それらの問題を解決するために、遺伝子相互作用関係データベースを構築するためのフレームワークを開発した。

図 29 に開発したフレームワークの概略を示す。

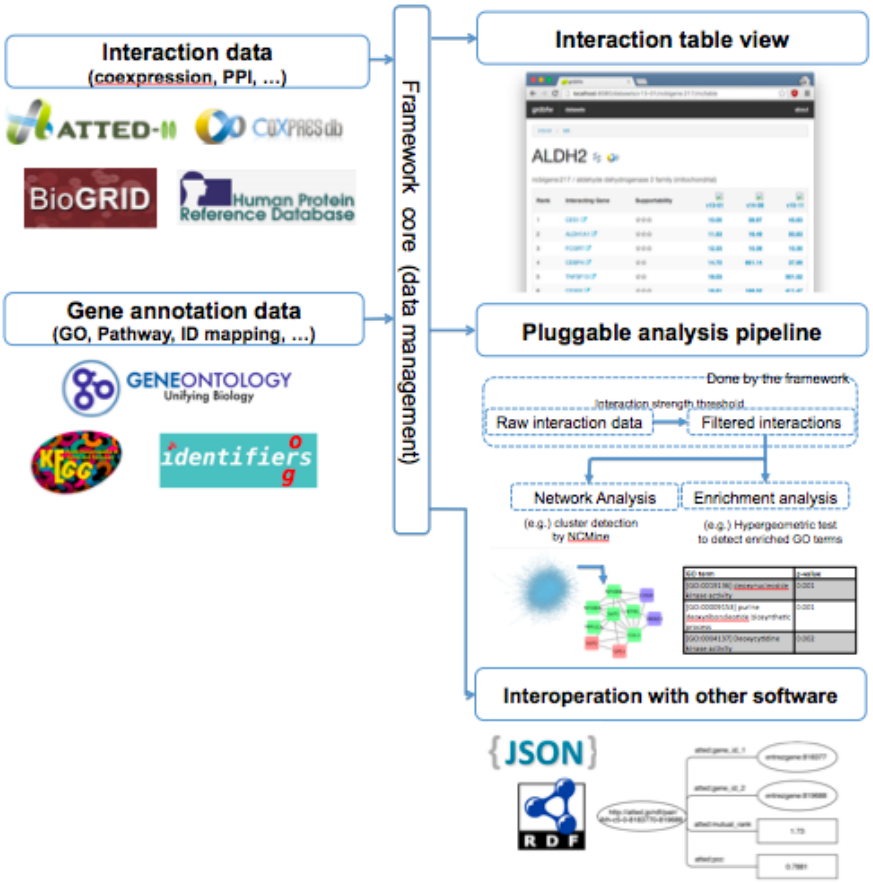



**図 29 PairDB フレームワーク概略図**

## 入力データ

フレームワークの入力として次の 2 つを与える。

- 共発現やタンパク質間相互作用などの遺伝子相互作用データ
- Gene Ontology (The Gene Ontology Consortium, 2015) や KEGG パスウェイ (Keshava Prasad *et al.*, 2009) などの遺伝子に対するアノテーションデータ

相互作用データとしては遺伝子共発現やタンパク質間相互作用データのような相互作用データが想定される。しかしながら、これに限らず、例えばタンパク質・リガンド結合データや、文献解析によって得られた遺伝子共起データなど、2 体の関係性データであればいかなるデー

タでも読み込むことが可能である。また、アノテーションデータとして、Gene Ontology や KEGG パスウェイなどを与えることができる。与えられたアノテーションデータは統計解析などに利用が可能である。

## フレームワークに実装されている機能や特徴

### 相互作用関係性テーブルの表示

読み込まれた相互作用データを表形式で提示する機能が実装されている。例えば、共発現データであれば、遺伝子を一つ選択すると、その遺伝子と共発現する遺伝子を共発現度が高い順にソートして表示する。このフレームワークには複数の相互作用データを読み込ませることが可能であり、「ホモログ遺伝子」などのような遺伝子間の関係性データも合わせて読み込ませることで、1 つのテーブルに複数の相互作用データをまとめて表示させることができる。これにより、複数の相互作用データの比較などを容易に行う事ができる。

### 相互作用データ解析

相互作用データに対し広く用いられる解析手法として、クラスタリングが挙げられる。これは相互作用の強さでフィルタリングを行い、強い相互作用を行う 2 遺伝子間にエッジが張られるようなネットワークを構築し、それに対してノードのクラスタリングを行う解析である。これにより、相互作用ネットワークから関係の深いノード群が抽出できると期待される。PairDB フレームワークには、標準で NCMine アルゴリズムと、それによって得られるモジュールに対し enrichment テストを行うルーチンが標準で組み込まれており、簡単なネットワーク解析を行うことが可能である。

相互作用データの解析手法としては、ネットワーククラスタリング以外にも様々な手法が考えられる。フレームワークの利用者が好みの解析手法を自由に追加することができるようにするため、PairDB フレームワークはその機能の一部を API として公開している。公開 API の仕様を満たすプラグインを作成することでフレームワークの機能を超えて好みの機能を追加することを容易に行うことができる。

## 他ソフトウェアとの連携

データベースは、単にデータを蓄積するだけではなく、蓄積されたデータを利用者に使っていただくことでその真価を発揮する。人間の利用者向けとして、相互作用関係性テーブルの表示や簡単なデータ解析手法を実装している。しかしながら人力によって全ての相互作用データを精査することは大変困難である。例えば、ヒト遺伝子は約 3 万とされているが、それらの間の全ての相互作用を検討する場合、3 万の 2 乗個分の相互作用データを確認する必要が発生する。そのような莫大なデータが計算機で自動処理させることができれば大変便利である。そのために、PairDB フレームワークでは他ソフトウェアとの連携が容易になる機能も実装している。それが REST API と RDF を通じたデータ提供である。

PairDB フレームワークは自動的にデータを JSON 形式に変換し、REST API を通じてデータを提供する機能を実装している。REST API によるデータ交換は Web 一般で広く用いられている手法である。また、近年、セマンティックウェブ技術が流行の兆しを見せている。特に生命科学分野においては、UniProt や PubChem といった大手データベースが自身のデータを RDF 形式で公開するなど、広く受け入れられつつある。PairDB フレームワークは自身のデータを自動的に RDF 形式に変換する機能を実装している。また、同時に SPARQL エンドポイントも公開することで、他データベースとのマッシュアップを容易にしてている。PairDB フレームワークの RDF 化機能は、ATTED-II・COXPRESdb の RDF 化に採用されているスキーマを用いて行われている。このスキーマの詳細は次節で説明する。

# 3.4 共発現データベース ATTED-II・COXPRESdb の RDF 化

今日、実験技術の向上などにより、遺伝子発現やタンパク質構造・ゲノム配列などの生命科学データの生産量は急速に増えている。また、生産されたデータを蓄積するデータベースも数多く作られている。 (Rigden *et al.*, 2016; Katayama *et al.*, 2013) そのような状況下で、個々の情報源からから提供されるデータを効率的に利用するために重要となるのが、データの統合である。セマンティックウェブ (Berners-Lee *et al.*, 2001) 技術は数ある解決策の中の一つであり、バイオインフォマティクス分野において、近年、人気が出てきている。セマンティックウェブでは、様々なデータを、統一された形式、すなわち、RDF 形式で表現する。

ATTED-II（http://atted.jp/） (Aoki *et al.*, 2016) と COXPRESdb（http://coxpresdb.jp） (Okamura *et al.*, 2015) は、それぞれ、植物・動物の遺伝子共発現データベースである。遺伝子共発現は、大量の DNA マイクロアレイや RNA-Seq の実験結果から計算され、遺伝子機能や遺伝子間の制御関係推定するために利用することができる。ATTED-II と COXPRESdb では従来、共発現データをテキストファイル・REST API を通じて提供していた。これに加え新たに RDF 形式での提供も行うようにした。

## セマンティックウェブと RDF・オントロジー

セマンティックウェブとは 2001 年頃、W3C の Tim Berners-Lee が提唱した概念である。これは、Web ページやデータベースに収録されているデータに対してアノテーション（メタデータ）を付加し、コンピュータ自身がデータの意味を解釈できるようにするといった試みである。コンピュータ自身がデータの意味を解釈できるようになると、大量のデータを自動的に処理できるようになるという利点がある。

このような概念を実現させるためには、統一されたデータモデルがあることが望ましい。このために作られたのが RDF（Resource description framework）と呼ばれるデータモデルである。RDF データモデルを用いると、すべてのデータは主語（Subject）・述語（Predicate）・目的語（Object）という 3 要素（トリプル）の集合、つまり、グラフモデルとして表現される。

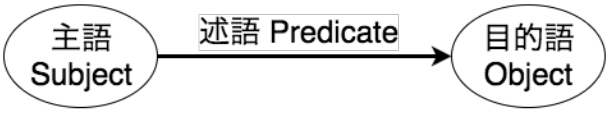



**図 30 トリプルの例**

例えば、「遺伝子 A が遺伝子 B を活性化する」というデータは次のようなトリプルで表すことができる。

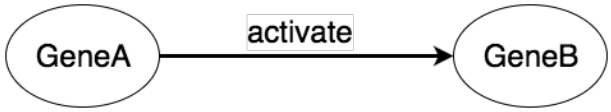



**図 31 「遺伝子 A が遺伝子 B を活性化する」というデータをトリプルで表した場合の例**

RDF データモデル（トリプル）を用いてデータを表現することは、ある物体間の関係性を示すグラフを生成・構築することとも言える。

RDF を用いてデータを表現することの利点の一つとして、データ構造の柔軟性が挙げられる。従来のデータベースは、伝統的に、関係データベースモデルを採用している。関係データベースモデルは表現したいデータを表形式に正規化し表現するモデル化の一つであり、一度、データを表形式にモデル化した後は、表に新しい列を追加することは大変困難であると言われている。対して、RDF で行われるグラフ形式の表現方法では、新しいデータが増えても、グラフのノード・エッジを増やすだけでデータの表現方法そのものを拡張することが可能である。
RDF データモデルを用いてデータを表現する場合、考慮するべき点として、オントロジーに関することが挙げられる。例えば、「"活性化する"という意味をどのような記号を用いて表すか」などのように、主に、述語に当たる部分にどのような記号を使うかということである。バイオインフォマティクス分野においてよく用いられるオントロジーは、主に、BioPortal（https://bioportal.bioontology.org/） (Whetzel *et al.*, 2011) と呼ばれるポータルサイトにまとめられている。

## 生命科学分野におけるセマンティックウェブ・RDF の利用

バイオインフォマティクス分野はセマンティックウェブ技術が比較的取り入れられつつある分野の一つである。その理由の一つとして RDF データモデルの柔軟性が挙げられる。生命科学分野では、実験手法の向上などにより、時間が進むにつれ、得られるデータの種類が増えてゆ

くということが起こり得る。その際に、新しい概念を付け加えやすい RDF データモデルがデータの表現形式として好まれる傾向にあると考えられる。

収録されているデータを RDF 形式で公開しているデータベースは数多く存在する。例えば、UniProt (The UniProt Consortium, 2015) である。UniProt タンパク質のデータベースであるが、生命科学分野においていち早くデータの RDF に取り組んできたデータベースの一つである。2016 年 11 月の時点では、自身のデータのほとんどを、合計 23000000000 個のトリプルとして表現し、公開している。他にも化合物データベースである PubChem (Kim *et al.*, 2016) などが収録データを RDF 形式で公開している。データベース運営者自身がデータの RDF 化を行っていなくても、第三者がデータの RDF 化を行っている場合もある。Bio2RDF (Belleau *et al.*, 2008) は、KEGG (Kanehisa *et al.*, 2014) や OMIM (Amberger *et al.*, 2015) など、RDF の提供を行っていないデータベースを RDF 化するプロジェクトである。

また、生命科学データの RDF 化やそれらの活用を目指す試みの一つとして、2008 年より、バイオサイエンスデータベースセンター（NBDC）・ライフサイエンスデータベースセンター（DBCLS）主導により BioHackathon（http://biohackathon.org/）が開催されている。これは生命情報科学分野の研究者・開発者を集め、集中的にセマンティックウェブ技術の基礎・応用アプリケーション開発を行う試みである。毎年、国内外を含め多くの開発者が集まり、ディスカッションを行っている。

このように、生命科学分野において、データを RDF 形式で表現することや、それを運用しようとする動きは急速に高まっている。

## 共発現データのモデル化

共発現データを RDF データモデルで表現するため、はじめに共発現データモデルスキーマを作成した。スキーマとは、ある種類のデータ（ここでは共発現データ）が RDF 上でどのようなトリプルとして表現されるかを示すものである。RDF データモデルは柔軟性に富んだデータモデルであるが、計算機による自動データ処理などを考えると、ある程度の制約が必要で、スキーマの生成・構築は、その制約の作成に当たると考えることができる。

調査の結果、遺伝子共発現やタンパク質相互作用のような 2 体関係を示す RDF データモデルは、まだ考案されていなかったため、新しいものを作成することとした。検討の結果、共発現データの RDF データモデル化として、次のようなデータモデルを考案した。スキーマ全体は参考資料 X に示す。図 32 は「遺伝子 818377 と遺伝子 819688 の共発現度が MR = 1.73, PCC

（Pearson’s correlation coefficient）= 0.7881 である」というデータを考案したデータモデルで示した図である。

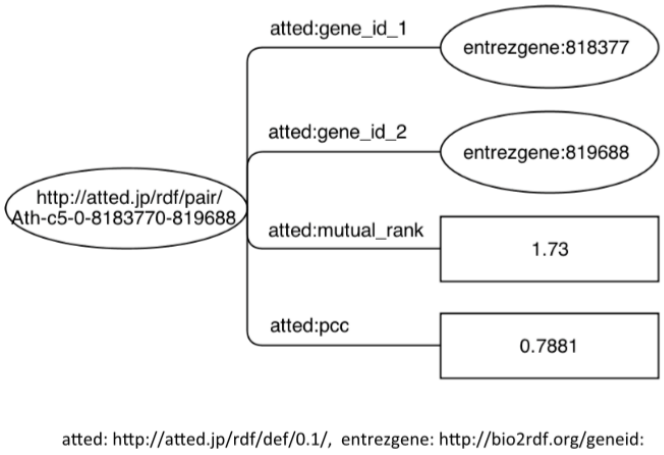



**図 32 遺伝子 818377 と遺伝子 819688 の共発現度を示すトリプル群**

木（グラフ）の根として、共発現する 2 遺伝子の識別子（名前）を含むノードを用意し、そのノードに共発現する 2 遺伝子と共発現度を示すノードを接続する仕様とした。根となるノードを用意しなくても遺伝子共発現をトリプルで表現することは可能であったが、実運用を考えた際、ユーザのクエリとして考えられるのは、

- ある遺伝子と共発現する遺伝子群が知りたい
- ある遺伝子 A とある遺伝子 B が共発現するのか知りたい

という 2 点であると考えられるため、2 点目を満足するには、根となるノードが存在したほうがグラフ検索に要する時間などの点で有利であるため、根を含める仕様とした。

遺伝子の ID（URI）としては、Bio2RDF (Belleau *et al.*, 2008) が定義するものを用いた。共発現度を示す指標である MR（Mutual Rank）や PCC（Pearson’s correlation coefficient）を表すオントロジーは存在していなかったため独自に定義を行った。それぞれ

- MR: http://atted.jp/rdf/def/0.1/mutual_rank
- PCC: http://atted.jp/rdf/def/0.1/pcc

を用いることとした。

## ATTED-II・COXPRESdb の RDF 化

前節で作成したスキーマを利用し、ATTED-II・COXPRESdb で提供している遺伝子共発現データの RDF 化を行った。ここでは、特に共発現度の高い遺伝子ペア（MR < 100）の共発現データに関して RDF 化を行った。共発現データのトリプル表現は、理論上、全ての遺伝子ペアに対して可能であるが、変換後の RDF データを保存しておくデータベースサーバの性能上の限界を鑑みた結果である。7 生物種の共発現データを RDF 化し、それぞれのトリプル数は表 12 の通りである。

変換後のデータは、データそのものを公開する他に、SPARQL エンドポイント（http://atted.jp/sparql, http://coxpresdb.jp/sparql）を通じて公開している。

**表 12 ATTED-II が取り扱う生物種と共発現データを RDF 化した際のトリプル数**

| 生物種 | 遺伝子数 | トリプル数 |
|---|---|---|
| **Arabidopsis** | 20836 | 7948911 |
| **Arabidopsis2** | 25610 | 8925810 |
| **Rice** | 20625 | 7809485 |
| **Soybean** | 15902 | 5626687 |
| **Maize** | 8397 | 3220164 |
| **Grape** | 8351 | 3456279 |

| | | |
|---|---|---|
| **Poplar** | 21909 | 6814735 |
| **Medicargo** | 4166 | 1869038 |
| **Total** | **125796** | **45671109** |

※Arabidopsis は測定プラットフォームの違いにより 2 種類存在する

## RDF 化されたデータの利用例

上記の SPARQL エンドポイントを通じて公開されたデータ、つまり、トリプルで構成されるグラフは、SPARQL 言語を使って検索することができる。図 33 に例を示す。

```
SELECT
  ?gene2 ?mr ?pcc
WHERE {
  <http://atted.jp/rdf/db/Ath.c5-0> atted:gene_pair ?pair .
  ?pair atted:gene_id_1 <http://bio2rdf.org/geneid:818377> .
  ?pair atted:gene_id_2 ?gene2 .
  ?pair atted:mutual_rank ?mr .
  ?pair atted:pcc ?pcc .
}
```

**図 33 遺伝子 818377 と共発現する遺伝子を検索する SPARQL クエリ例**

図 33 は遺伝子 818377 と共発現する遺伝子と共発現度を得るための SPQRQL クエリである。SPARQL では、基本的に、WHERE 節でマッチさせたいトリプルの形を指定することによりクエリを行う。

共発現データの RDF 化を行い SPARQL エンドポイント通じた検索が可能になる利点として、他のデータソースとシームレスに連携可能であるということが挙げられる。タンパク質のデータベースである UniProt などいくつかのデータベースは積極的に収録データの RDF 化を推進し、また、SPARQL エンドポイントを公開している。SPARQL の機能として Federated Query と呼ばれるものがある。図 34 は Federated Query の概念を示す。

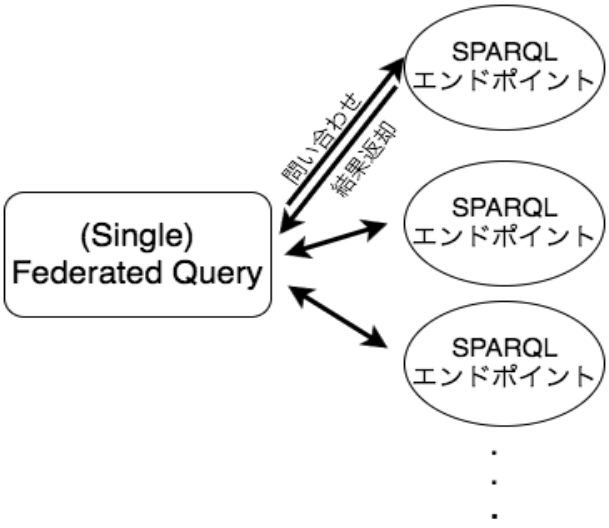



**図 34 Federated SPARQL query の概念図**

Federated Query を用いることで、複数のデータソースに対しクエリを発行しその結果をまとめることができる。この場合のデータソースには、ローカルのデータベースに限らず、遠隔地にあるデータベースに収録されているデータを含めることが可能であり、より効率的なデータ統合を行うことができる。

図 35 に Federated Query を用いて、ATTED-II・COXPRESdb の共発現データと UniProt のタンパク質データをシームレスに統合する SPARQL クエリの例を示す。

```
SELECT
   ?go_id ?go_description COUNT(*)
WHERE {
   # selects genes with coexpress with 'geneid:818377'
   <http://atted.jp/rdf/db/Ath.c5-0> atted:gene_pair ?pair .
   ?pair atted:gene_id_1 <http://bio2rdf.org/geneid::818377> .
   ?pair atted:gene_id_2 ?gene2 .
   ?pair atted:mutual_rank ?mr .
   ?pair atted:pcc ?pcc .

   # analyzes GO term distribution of coexpressing genes
   SERVICE <http://go.bio2rdf.org/sparql> {
      ?gene2 <http://bio2rdf.org/geneid_vocabulary:process> ?go_id .
      ?go_id rdfs:label ?go_description
   }
}
GROUP BY ?go_id ?go_description
ORDER BY DESC(COUNT(*))
```

**図 35 遺伝子 818377 (AT2G37990 ribosome biogenesis regulatory protein (RRS1) family protein)と共発現する遺伝子に付与されている GO タームの解析を行うクエリ例**

| go_id | go_description |
|---|---|
| http://bio2rdf.org/go:0009220 | pyrimidine ribonucleotide biosynthetic process [go:0009220] |
| http://bio2rdf.org/go:0008150 | biological process [go:0008150] |
| http://bio2rdf.org/go:0001510 | RNA methylation [go:0001510] |
| http://bio2rdf.org/go:0006606 | protein import into nucleus [go:0006606] |
| http://bio2rdf.org/go:0009909 | regulation of flower development [go:0009909] |
| http://bio2rdf.org/go:0009640 | photomorphogenesis [go:0009640] |

**図 36 SPARQL クエリの実行結果例**

図 35 は ATTED-II から遺伝子 818377（AT2G37990 ribosome biogenesis regulatory protein (RRS1) family protein）と共発現する遺伝子群を抽出し、抽出された遺伝子群を Bio2RDF の SPARQL エンドポイント（http://go.bio2rdf.org/sparql）へ渡すことで、共発現する遺伝子それぞれ GO（Gene Ontology）タームを得る。そしてそれを集計している。このように ATTED-II と Bio2RDF の 2 つのデータ・ソースを使い、（i）共発現する遺伝子の取得、（ii）GO タームの関連付け、（ii）GO タームの集計を 1 個の SPARQL クエリで実現している。

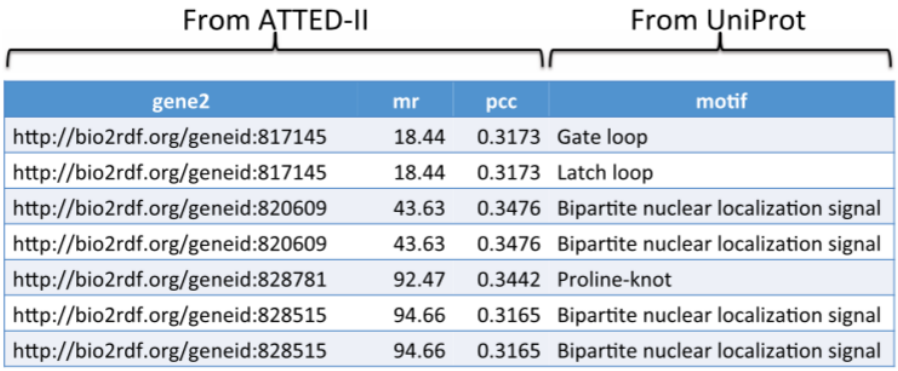
図 37 は 2 個目の例である。

```
SELECT
   ?gene2 ?mr ?pcc ?motif
WHERE {
   # selects genes with coexpress with 'geneid:81864' (ATTED-II)
   <http://atted.jp/rdf/db/Ath.c5-0> atted:gene_pair ?pair .
   ?pair atted:gene_id_1 <http://bio2rdf.org/geneid::81864> .
   ?pair atted:gene_id_2 ?gene2 .
   ?pair atted:mutual_rank ?mr .
   ?pair atted:pcc ?pcc .

   # converts URI for genes (Bio2RDF)
   SERVICE <http://uniprot.bio2rdf.org/sparql> {
       ?gene2 owl:sameAs ?uniprot_gene_uri
   }

   # selects motifs related to coexpressing genes (UniProt)
   SERVICE <http://beta.sparql.uniprot.org/sparql> {
       ?protein a uniprot:Protein .
       ?protein rdfs:seeAlso ?uniprot_gene_uri .
       ?protein uniprot:annotation ?annotation .
       ?annotation a uniprot:Motif_Annotation .
       ?annotation rdfs:comment ?motif
   }
}
```

**図 37 ATTED-II より遺伝子 81864（EVR3 exudative vitreoretinopathy 3）と共発現する遺伝子を取得し、UniProt よりそれらの遺伝子のモチーフを取得する SPARQL クエリ例**




| gene2 | mr | pcc | motif |
|---|---|---|---|
| http://bio2rdf.org/geneid:817145 | 18.44 | 0.3173 | Gate loop |
| http://bio2rdf.org/geneid:817145 | 18.44 | 0.3173 | Latch loop |
| http://bio2rdf.org/geneid:820609 | 43.63 | 0.3476 | Bipartite nuclear localization signal |
| http://bio2rdf.org/geneid:820609 | 43.63 | 0.3476 | Bipartite nuclear localization signal |
| http://bio2rdf.org/geneid:828781 | 92.47 | 0.3442 | Proline-knot |
| http://bio2rdf.org/geneid:828515 | 94.66 | 0.3165 | Bipartite nuclear localization signal |
| http://bio2rdf.org/geneid:828515 | 94.66 | 0.3165 | Bipartite nuclear localization signal |

**図 38 SPARQL クエリの実行結果例**

図 37 は、ATTED-II より遺伝子 81864（EVR3 exudative vitreoretinopathy 3）と共発現する遺伝子を取得し、UniProt よりそれらの遺伝子の配列モチーフを取得する SPARQL クエリ例である。この例でも、（i）ATTED-II からある遺伝子と共発現する遺伝子を取得、（ii）Bio2RDF を用いて遺伝子 ID の変換を行う、（iii）UniProt を用いて共発現する遺伝子の配列モチーフを取得するといった、複数データソースの統合を 1 クエリで実現している。

# 3.5 コホート向けメタボロームデータの解析プラットフォーム MetaboloView の開発

メタボロミクスは、様々な細胞プロセセスによって生産される代謝産物（低分子）を質量分析器を用いて、網羅的に同定・特徴づけを行う研究の一つである。メタボロミクス研究を行うことで、代謝パスウェイの動きの変化を検出される代謝物の差として検知することができる。そのため、表現型的な変化が起こる前にその変化を予測すること、つまり、疾患の進行などの検知に用いることができる。そのため、今日、メタボロミクス研究は様々な分野で重要な役割を担っている。こうした背景から、MZmine (Pluskal *et al.*, 2010) や MetDAT (Biswas *et al.*, 2010) などの、メタボロミクス研究を支援するためのソフトウェアも数多く開発されてきた。

最近では、メタボロミクス分野でコホート研究が行われ始めている。例えば、HUSERMET プロジェクト (Dunn *et al.*, 2011) や鶴岡メタボロームコホート研究 (Harada *et al.*, 2016) などである。コホート研究で数多くのメタボローム解析サンプルを収集することで、疾患の特徴づけの際に鍵となる因子（疾患マーカー）の同定や創薬ターゲットの探索などへの応用が期待される。しかし、データの蓄積が進む一方、蓄積されるデータの解析基盤が十分とは言えない。MZmine や MetDAT はは単一、または数個のサンプルの解析には適しているが、コホート研究といった数百、数千サンプルのデータ解析を解析するという点では効率的に動作しない。また、大規模コホートデータに適した解析基盤は今のところ存在しない。この問題点を解決するため、MetaboloView と呼ばれる大規模コホートデータ向けのデータ解析基盤を構築した。

## MetaboloView によるメタボロームデータ解析

図 39 に MetaboloView を用いたメタボロームデータの解析フローを示す。

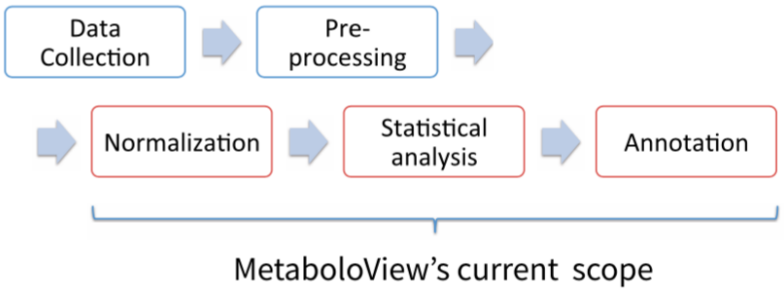



**図 39 MetaboloView を用いたデータ解析フロー**

### 3. 大規模データ解析補助アプリケーションの開発

データの収集・前処理、行った後、データの正規化・統計解析・化合物に対するアノテーションを行う。MetaboloView が対象とするのは、主にメタボロームデータの正規化・統計解析・アノテーションの部分である。

図 40 に MetaboloView のメイン画面を示す。

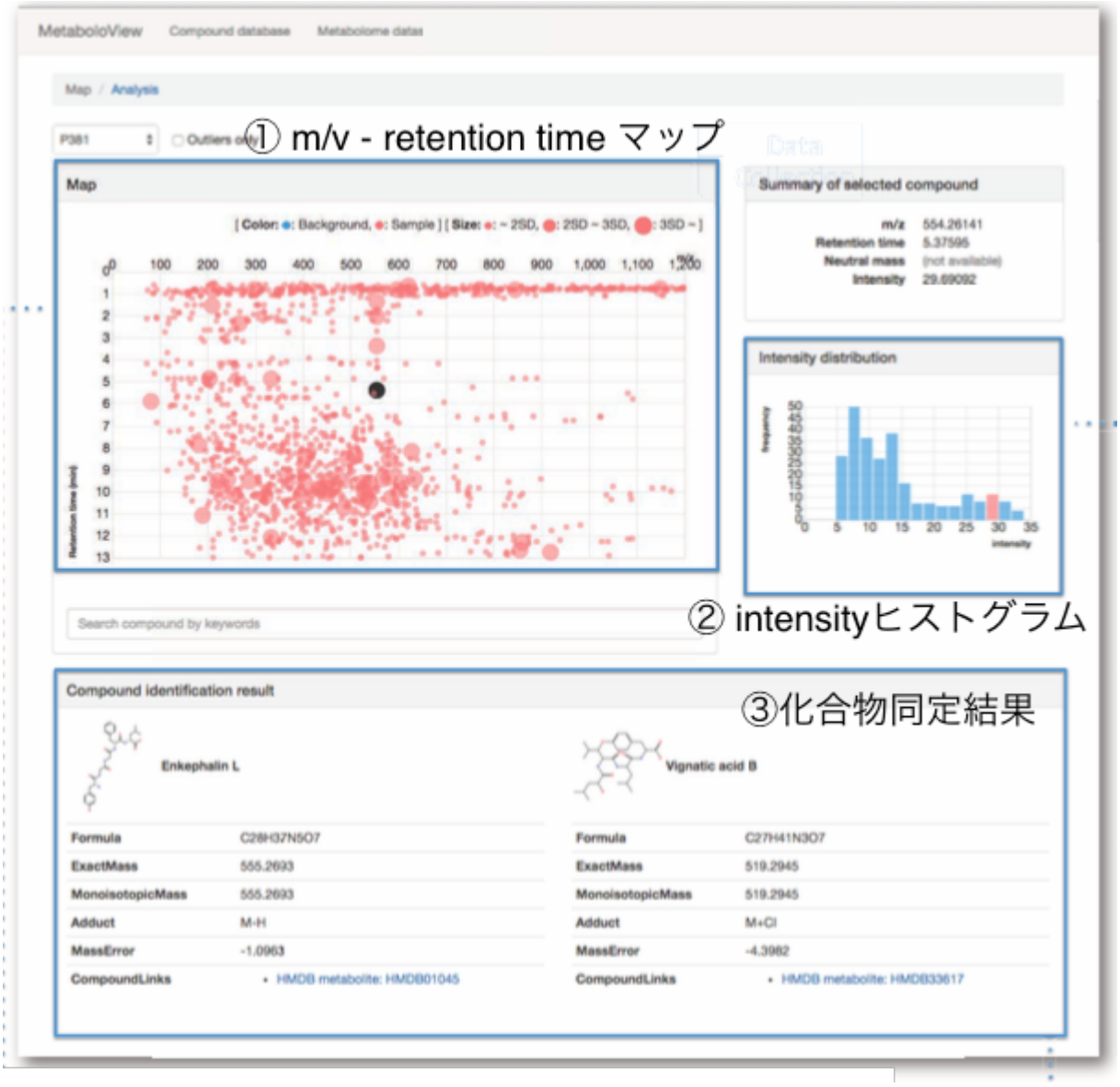



**図 40 MetaboloView メイン画面**

### 3. 大規模データ解析補助アプリケーションの開発

メイン画面では大きく分けて 3 個の機能（m/z-retention time マップ、intensity ヒストグラム、化合物同定結果表示）が利用できる。

## m/z - retention time マップ

質量分析器で検出された化合物は、それぞれ、質量（m/z 値）と保持時間（retention time）の情報を持っている。この値をそれぞれ横軸・縦軸にとったプロットのことを MetaboloView では m/z - retention time マップと呼んでいる。このプロットの 1 点が 1 低分子に対応する。このマップを見ることで、検出された低分子全体の傾向を読み取ることが可能である。

また、このプロットでは点の大きさが、質量分析器で検出された時の強度（intensity 値）を示す。ただし、実際には intensity の値そのものが示されているわけではない。大量のメタボロームサンプルがある場合、intensity 値の分布を考えることが可能であり、点の大きさは全サンプルにおける特定化合物の intensity 値の平均値からのずれを示している。このように表示することにより、化合物の intensity と、疾患のような特定の表現型の関係性を検討しやすくなる。MetaboloView にはずれが大きい化合物のみを表示するモードも実装されている。

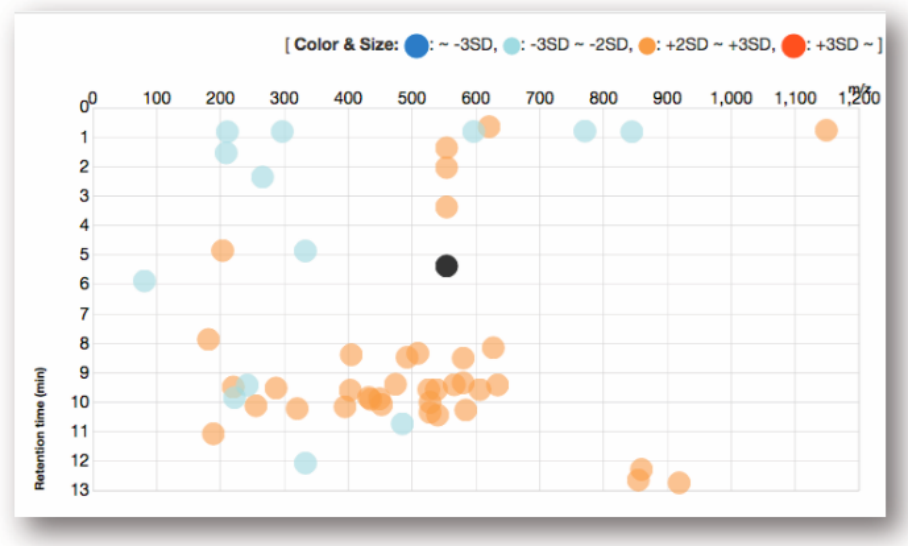



**図 41 ずれが大きい化合物のみを表示した場合**

### Intensity ヒストグラム

これは intensity の分布をヒストグラムとして表示する機能である。ヒストグラム中のバーを選択することで、そのバーが表す intensity の範囲に含まれるサンプルを選択することが可能である。この機能も化合物の intensity 値と疾患の関係性の解析を容易に進めるために実装した。

### 化合物同定結果の表示

質量分析器で検出された低分子が既知の化合物であるかどうか知ることは、サンプル解析において重要である。MetaboloView では化合物データベースである Human Metabolome database（HMDB） (Wishart *et al.*, 2013) と LIPID MAPS (Sud *et al.*, 2007) を取り込み、検出された低分子の m/z 値で検索を行うことにより、検出された化合物の同定を行っている。HMDB・LIPID MAPS 以外にも化合物データベースの取り込みは容易に行うことができる。

### 統計解析

MetaboloView は種々の統計解析手法も実装しており、MetaboloView に読み込まれたデータに対して自動的に様々な統計解析を行う。例えば図 42 は検出された低分子の intensity 値を用いて測定サンプルをクラスタリングした結果である。

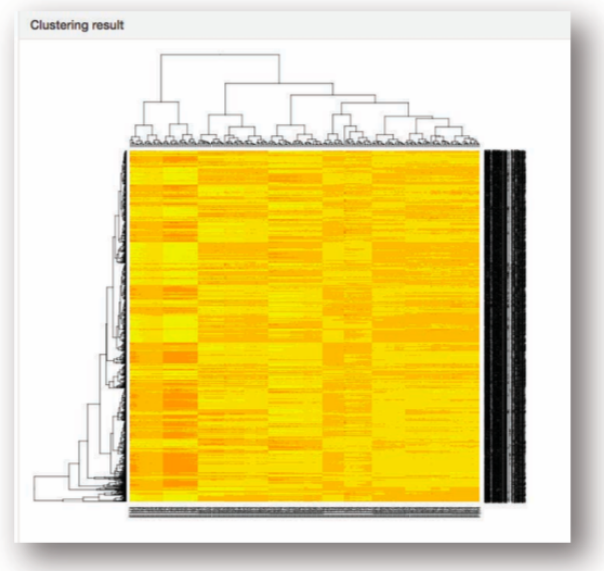



**図 42 測定サンプルのクラスタリング解析を行う例**

### 3. 大規模データ解析補助アプリケーションの開発

MetaboloView は大規模コホートデータの取り扱いに焦点を当てた解析基盤であるため、大規模データ向けの統計解析手法も実装している。例えば、一回の質量分析器の運転で計測できるサンプル数は限られているため、コホートのように大量にサンプルを読まなければならない場合、複数回、質量分析器を運転する必要がある。その際、各運転間で、首尾一貫しているデータが出ているかどうかなどを確認する必要がある。図 43 は質量分析器の運転で得られたデータの相関を計算するような解析の結果である。このような解析を組み込むことで、特に大規模メタボロミクスデータの質に関する検討を行うことができるようになる。

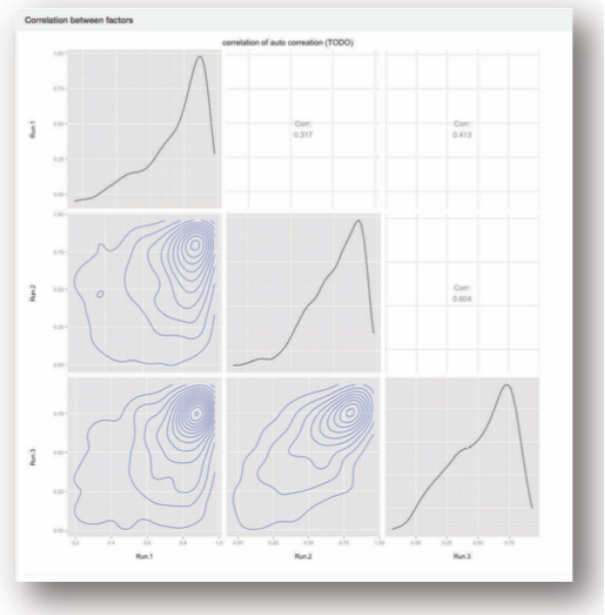



**図 43 データの一貫性の検討を行うための解析結果の例**

## MetaboloView の実装

MetaboloView は Java 言語と Javascript 言語を用いて実装されている。また、MetaboloView の機能の一部は Java の interface を通じて外部へ公開されている。MetaboloView 本体はメタボロームデータの管理と基本的な可視化を備えており、それ以外の部分は公開 API を通じて拡張が可能である。例えば、前節で上げた統計解析手法は MetaboloView の公開 API を用いて MetaboloView 本体へ組み込まれている。このように、任意の処理がプラグインという形で追加可能であるため、MetaboloView を目的に合う形に変化させることも容易に行うことができる。

## 3.6 小括

近年、ゲノム配列・タンパク質構造情報など、あらゆる生命関連データが急速に増えている。増えつつけるデータを扱いやすい形で管理・運用し、さらに、そのデータの山から生命科学的に意味のあることを見つけ出す手助けを行うことは情報科学的に大変意義のあることである。筆者は、遺伝子関係性データベースフレームワーク PairDB の構築・共発現データベース ATTED-II・COXPRESdb の RDF 化を通じ、データの効率的な運用を行う基盤を構築し、また、遺伝子相互作用ネットワーク解析アルゴリズム NCMine の考案・コホート向けメタボローム データ解析基盤 MetaboloView の実装を通じ、データの山から意味のあるデータを見出すための基盤を実装してきた。

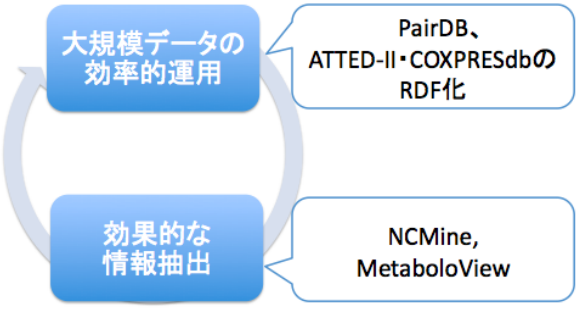



**図 44 効率的な大規模データ運用・解析**

近年では、生命科学分野に限らず、様々な分野で「ビッグデータ」という言葉が使われる。私は、大規模データの運用・解析の歯車を効率よく回すことが、データの山から新しいデータを引き出す上で大変重要であると考えている。

# 総括

第 1 章では相互作用ネットワークより、機能モジュールを検出する手法の開発に関して述べた。これまでに計算機的に相互作用ネットワーク中から機能モジュールを抽出・予測する研究が行われてきた。しかし既存の手法には抽出するモジュールの構造に関する問題や手法の効率に関する問題がある。これらの問題点を解決し、さらに、機能モジュールのダイナミクスに関する知見を得ることを目的とし、抽出手法の開発を行った。

既存研究では完全グラフの探索により、機能モジュールの抽出やタンパク質複合体の予測に成功したものがある。しかし、機能モジュールを構成するタンパク質間に全対全の相互作用がある訳ではなく、完全グラフという制約は生物学的な相互作用ネットワークに対し、厳しすぎる制約である。そこでその制約を弱めることで、より実際の構造に近い機能モジュール抽出を行えるのではないかと考えた。ここで完全グラフから何本かエッジが欠けたグラフを near-clique と定義する。提案手法は near-clique の探索を行い、発見された near-clique にさらに処理を加えたものを機能モジュールとして出力するものである。また、過去の研究により機能モジュールには core-peripheral 構造と呼ばれる構造的特徴があることが分かっている。それは図 2 に示されるように、モジュール内のタンパク質はモジュールの中心的役割を果たす core タンパク質と core のうちのいくつかと相互作用する peripheral タンパク質の 2 種類に分類することができる。peripheral は動的に変化するため、この構造的特徴は生命機能のダイナミクスを解明する上で重要な情報源である。これまでに考案されてきた機能モジュール抽出を行うための手法は主に相互作用ネットワーク内から特に密に接続されたサブネットワークの抽出であった。それはある機能モジュールに参加するタンパク質間同士にはより強い相互作用があると考えられるためである。過去にはその仮定を利用し、多くの手法が提案されてきたが、モジュールの構造的特徴を捉えることのできる手法は存在しない。提案手法は near-clique の探索後、ノードの重なりによるクラスタリングを行い near-clique 群の階層関係を得る。そしてその結果を最終的な機能モジュール群として出力する。この near-clique のクラスタリングによりモジュールの core-peripheral 構造を明らかにした。

既存手法は、大規模な相互作用ネットワークからモジュール抽出を行う際、実行時間や計算量といった探索効率の面で問題が見られる。例えば、ヒトのタンパク質間相互作用ネットワークはおよそ 3 万のタンパク質を含んでいる巨大なネットワークである。そのため効率よく機能モジュールを抽出する必要がある。提案手法ではネットワーク中のノードの重要度付け手法を利用し、重要度の高さに従ってネットワークを走査することで効率的・高速なモジュール抽出を実現した。

## 総括

第 2 章では複数の分子ネットワークの統合を試みた。これまでに生命を表現する分子ネットワークとしていくつかの種類のネットワークが考案されてきている。例えば、タンパク質間相互作用ネットワーク・遺伝子共発現ネットワーク・シグナル伝達ネットワークと呼ばれるものである。それぞれのネットワークを作成するために必要な実験や計算が異なり、また、異なった特徴を持っている。しかしながら、計算方法などデータ元の違いに関して比較的よく知られている一方、ネットワークとしての差異や、それぞれの利点・欠点に関する深い検討はなされていない。第 2 章ではタンパク質間相互作用ネットワークと遺伝子共発現ネットワークを例に挙げ、これらの比較検討を行った。また、続いてタンパク質間相互作用ネットワークと遺伝子共発現ネットワークの統合解析に関する検討を行った。

統合によって得られたネットワークに対し、第 1 章で構築した機能モジュール抽出手法を適用すると、抽出されたクラスタの中に、実際のパスウェイを再現していると思われるクラスタを見出すことができた。このクラスタには、タンパク質相互作用に対応するエッジと遺伝子興発現に対応するエッジ両方が含まれており、統合ネットワークからでなければ見出せないクラスタであった。

第 3 章では、大規模生命関連データの解析補助を行うアプリケーションの開発に関して述べた。実験データの生産スループットの上昇は、生物に関する深い知見をもたらすが、同時に、生産されるデータの管理や解析方法といった点で問題を引き起こす。例えば、一度の実験で生産されるデータの量はテラバイトに及ぶこともあり、人の手で扱えるものではないという場合も多々発生している。そのような場において、生産されるデータを整理された形で蓄積し、さらに蓄積された膨大なデータの山から生物学的に意味のあることを見出すための基盤やアルゴリズムを開発することは情報科学分野として大変重要であり、同時に、意義のあることである。

第 1 章で述べた遺伝子ネットワークより機能モジュールを抽出する手法は、アプリケーションとしての実装を行い、フリーソフトウェアや Web サービスとして公開している。また、筆者は遺伝子ネットワークに対象を絞らず、様々な種類の生命関連データを保持・解析するための基盤アプリケーションを作成している。例えば、大規模メタボロミクスデータの解析・可視化基盤の構築や、大規模遺伝子関係性データベース構築基盤の開発・遺伝子相互作用データの RDF 表現の作成なども行っている。これら全てを通し、データの山から意味のあるデータを見出すことを目標としている。近年、生命科学分野に限らず様々な分野で「ビッグデータ」という言葉が用いられる。筆者は、大規模データ運用・解析の歯車を効率よく回すことが、データの山から新しいデータを引き出す上で大変重要であると考えている。

本論文を通し、筆者は情報科学的視点から、相互作用ネットワークや大規模生命関連データに関するデータ管理フレームワークや解析アルゴリズムを開発し、データ解析支援という面において生物学分野に対して一定の貢献がきたと考えている。これからもこのような貢献を続け

ていきたいと考えている。しかしながら、本論文では主に情報科学的視点を全面に生物学分野に対する支援の検討を行ってきた。より深い貢献を行うためには、生物学的分野の方々との協調が必要である。今後は、他分野との交流を広く行い、問題を発見し、発見した問題を筆者の専門分野である情報科学的なアプローチを元に解決するなど、多面的な問題解決を行っていきたい。

## 総括

# 引用文献

Adamcsek,B., Palla,G., Farkas,I.J., Derényi,I., and Vicsek,T. (2006) CFinder: Locating cliques and overlapping modules in biological networks. *Bioinformatics*, **22**, 1021–1023.

Addison,C.L., Daniel,T.O., Burdick,M.D., Liu,H., Ehlert,J.E., Xue,Y.Y., Buechi,L., Walz,A., Richmond,A., and Strieter,R.M. (2000) The CXC Chemokine Receptor 2, CXCR2, Is the Putative Receptor for ELR+ CXC Chemokine-Induced Angiogenic Activity. *J. Immunol.*, **165**, 5269–5277.

Albert,R. (2005) Scale-free networks in cell biology. *J. Cell Sci.*, **118**, 4947–4957.

Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F., and Hamosh,A. (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.

Andersen,R., Gharan,S.O., Peres,Y., and Trevisan,L. (2016) Almost Optimal Local Graph Clustering Using Evolving Sets. *J. ACM*, **63**, 1–31.

Aoki,Y., Okamura,Y., Tadaka,S., Kinoshita,K., and Obayashi,T. (2016) ATTED-II in 2016: A plant coexpression database towards lineage-specific coexpression. *Plant Cell Physiol.*, **57**, e5.

Bader,G. and Hogue,C. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.

Barabási,A.-L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.

Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M., Yefanov,A., Lee,H., Zhang,N., Robertson,C.L., Serova,N., Davis,S., and Soboleva,A. (2013) NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Res.*, **41**.

Belleau,F., Nolin,M.A., Tourigny,N., Rigault,P., and Morissette,J. (2008) Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inform.*, **41**, 706–716.

Berggård,T., Linse,S., and James,P. (2007) Methods for the detection and analysis of protein-protein interactions. *Proteomics*, **7**, 2833–2842.

Berners-Lee,T., Hendler,J., and Lassila,O. (2001) The Semantic Web A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities.

## 引用文献


Biswas,A., Mynampati,K.C., Umashankar,S., Reuben,S., Parab,G., Rao,R., Kannan,V.S., and Swarup,S. (2010) MetDAT: a modular and workflow-based free online pipeline for mass spectrometry data processing, analysis and interpretation. *Bioinformatics*, **26**, 2639–40.

Datt,S. (2011) The information explosion: Trends in Technology Review. *J. Gov. Financ. Manag.*, **60**, 46–51.

Derényi,I., Palla,G., and Vicsek,T. (2005) Clique percolation in random networks. *Phys. Rev. Lett.*, **94**.

Dunn,W.B., Broadhurst,D., Begley,P., Zelena,E., Francis-McIntyre,S., Anderson,N., Brown,M., Knowles,J.D., Halsall,A., Haselden,J.N., Nicholls,A.W., Wilson,I.D., Kell,D.B., Goodacre,R., and Human Serum Metabolome,H.C. (2011) Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat. Protoc.*, **6**, 1060–1083.

Gavin,A.-C., Aloy,P., Grandi,P., Krause,R., Boesche,M., Marzioch,M., Rau,C., Jensen,L.J., Bastuck,S., Dümpelfeld,B., Edelmann,A., Heurtier,M.-A., Hoffman,V., Hoefert,C., Klein,K., Hudak,M., Michon,A.-M., Schelder,M., Schirle,M., Remor,M., Rudi,T., Hooper,S., Bauer,A., Bouwmeester,T., Casari,G., Drewes,G., Neubauer,G., Rick,J.M., Kuster,B., Bork,P., Russell,R.B., and Superti-Furga,G. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.

Girvan,M. and Newman,M.E.J. (2002) Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 7821–6.

Harada,S., Takebayashi,T., Kurihara,A., Akiyama,M., Suzuki,A., Hatakeyama,Y., Sugiyama,D., Kuwabara,K., Takeuchi,A., Okamura,T., Nishiwaki,Y., Tanaka,T., Hirayama,A., Sugimoto,M., Soga,T., and Tomita,M. (2016) Metabolomic profiling reveals novel biomarkers of alcohol intake and alcohol-induced liver injury in community-dwelling men. *Environ. Health Prev. Med.*, **21**, 18–26.

Kanehisa,M., Goto,S., Sato,Y., Kawashima,M., Furumichi,M., and Tanabe,M. (2014) Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Res.*, **42**.

Katayama,T., Wilkinson,M.D., Micklem,G., Kawashima,S., Yamaguchi,A., Nakao,M., Yamamoto,Y., Okamoto,S., Oouchida,K., Chun,H.-W., Aerts,J., Afzal,H., Antezana,E., Arakawa,K., Aranda,B., Belleau,F., Bolleman,J., Bonnal,R.J.P., Chapman,B., Cock,P.J.A., Eriksson,T., Gordon,P.M.K., Goto,N., Hayashi,K., Horn,H., Ishiwata,R., Kaminuma,E., Kasprzyk,A., Kawaji,H., Kido,N., Kim,Y.J., Kinjo,A.R., Konishi,F., Kwon,K.-H., Labarga,A., Lamprecht,A.-L., Lin,Y., Lindenbaum,P., McCarthy,L., Morita,H., Murakami,K., Nagao,K., Nishida,K., Nishimura,K., Nishizawa,T., Ogishima,S., Ono,K., Oshita,K., Park,K.-J.,

Prins,P., Saito,T.L., Samwald,M., Satagopam,V.P., Shigemoto,Y., Smith,R., Splendiani,A., Sugawara,H., Taylor,J., Vos,R.A., Withers,D., Yamasaki,C., Zmasek,C.M., Kawamoto,S., Okubo,K., Asai,K., and Takagi,T. (2013) The 3rd DBCLS BioHackathon: improving life science data integration with Semantic Web technologies. *J. Biomed. Semantics*, **4**, 1–17.

Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A., Balakrishnan,L., Marimuthu,A., Banerjee,S., Somanathan,D.S., Sebastian,A., Rani,S., Ray,S., Harrys Kishore,C.J., Kanth,S., Ahmed,M., Kashyap,M.K., Mohmood,R., Ramachandra,Y.L., Krishna,V., Rahiman,B.A., Mohan,S., Ranganathan,P., Ramabadran,S., Chaerkady,R., and Pandey,A. (2009) Human Protein Reference Database--2009 update. *Nucleic Acids Res.*, **37**, D767–D772.

Khanin,R. and Wit,E. (2006) How Scale-Free Are Biological Networks. *J. Comput. Biol.*, **13**, 810–818.

Kim,S., Thiessen,P.A., Bolton,E.E., Chen,J., Fu,G., Gindulyte,A., Han,L., He,J., He,S., Shoemaker,B.A., Wang,J., Yu,B., Zhang,J., and Bryant,S.H. (2016) PubChem substance and compound databases. *Nucleic Acids Res.*, **44**, D1202–D1213.

Kuhner,S., van Noort,V., Betts,M.J., Leo-Macias,A., Batisse,C., Rode,M., Yamada,T., Maier,T., Bader,S., Beltran-Alvarez,P., Castano-Diez,D., Chen,W.-H., Devos,D., Guell,M., Norambuena,T., Racke,I., Rybin,V., Schmidt,A., Yus,E., Aebersold,R., Herrmann,R., Bottcher,B., Frangakis,A.S., Russell,R.B., Serrano,L., Bork,P., and Gavin,A.-C. (2009) Proteome Organization in a Genome-Reduced Bacterium. *Science.*, **326**, 1235–1240.

Latanya Sweeney,P.D. (2006) Information explosion. *Data Strateg.*, 16–18.

Li,X.-L., Tan,S.-H., Foo,C.-S., and Ng,S.-K. (2005) Interaction graph mining for protein complexes using local clique merging. *Genome Inform.*, **16**, 260–269.

Liu,G., Wong,L., and Chua,H.N. (2009) Complex discovery from weighted PPI networks. *Bioinformatics*, **25**, 1891–1897.

Lotia,S., Montojo,J., Dong,Y., Bader,G.D., and Pico,A.R. (2013) Cytoscape app store. *Bioinformatics*, **29**, 1350–1351.

Narushima,Y., Kozuka-Hata,H., Tsumoto,K., Inoue,J.-I., and Oyama,M. (2016) Quantitative phosphoproteomics-based molecular network description for high-resolution kinase-substrate interactome analysis. *Bioinformatics*, **32**, 2083–8.

Obayashi,T. and Kinoshita,K. (2009) Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. *DNA Res.*, **16**, 249–260.

Okamura,Y., Aoki,Y., Obayashi,T., Tadaka,S., Ito,S., Narise,T., and Kinoshita,K. (2015) COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. *Nucleic Acids Res.*, **43**, D82–D86.

Oltvai,Z.N. and Barabási,A. (2002) Life's Complexity Pyramid. *Science.*, **298**, 763–764.

Oughtred,R., Chatr-Aryamontri,A., Breitkreutz,B.J., Chang,C.S., Rust,J.M., Theesfeld,C.L., Heinicke,S., Breitkreutz,A., Chen,D., Hirschman,J., Kolas,N., Livstone,M.S., Nixon,J., O'Donnell,L., Ramage,L., Winter,A., Reguly,T., Sellam,A., Stark,C., Boucher,L., Dolinski,K., and Tyers,M. (2016) Biogrid: A resource for studying biological interactions in yeast. *Cold Spring Harb. Protoc.*, **2016**, 29–34.

Palla,G., Derényi,I., Farkas,I., and Vicsek,T. (2005) Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, **435**, 814–8.

Parkinson,H., Sarkans,U., Shojatalab,M., Abeygunawardena,N., Contrino,S., Coulson,R., Farne, a, Lara,G.G., Holloway,E., Kapushesky,M., Lilja,P., Mukherjee,G., Oezcimen, a, Rayner,T., Rocca-Serra,P., Sharma, a, Sansone,S., and Brazma, a (2005) ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **33**, D553–D555.

Pluskal,T., Castillo,S., Villar-Briones,A., and Oresic,M. (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, **11**, 395.

Qi,Y., Balem,F., Faloutsos,C., Klein-Seetharaman,J., and Bar-Joseph,Z. (2008) Protein complex identification by supervised graph local clustering. *Bioinformatics*, **24**.

Rao,V.S., Srinivas,K., Sujini,G.N., and Kumar,G.N.S. (2014) Protein-Protein Interaction Detection: Methods and Analysis. *Int. J. Proteomics*, **2014**, 1–12.

Rigden,D.J., Fernández-Suárez,X.M., and Galperin,M.Y. (2016) The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection. *Nucleic Acids Res.*, **44**, D1-6.

Rivera,C.G., Vakil,R., and Bader,J.S. (2010) NeMo: Network Module identification in Cytoscape. *BMC Bioinformatics*, **11 Suppl 1**, S61.

Rives,A.W. and Galitski,T. (2003) Modular organization of cellular networks. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 1128–1133.

## 引用文献


Seok-Joo Lee, Sungmin Kang, Hongyeon Kim, and Jun-Ki Min (2016) An efficient parallel graph clustering technique using Pregel. In, *2016 International Conference on Big Data and Smart Computing (BigComp)*. IEEE, pp. 370–373.

Shi,L., Lei,X., and Zhang,A. (2011) Protein complex detection with semi-supervised learning in protein interaction networks. *Proteome Sci.*, **9**, S5.

Smoot,M.E., Ono,K., Ruscheinski,J., Wang,P.L., and Ideker,T. (2011) Cytoscape 2.8: New features for data integration and network visualization. *Bioinformatics*, **27**, 431–432.

Spirin,V. and Mirny,L.A. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 12123–12128.

Sud,M., Fahy,E., Cotter,D., Brown,A., Dennis,E.A., Glass,C.K., Merrill,A.H., Murphy,R.C., Raetz,C.R.H., Russell,D.W., and Subramaniam,S. (2007) LMSD: LIPID MAPS structure database. *Nucleic Acids Res.*, **35**.

Szklarczyk,D., Franceschini,A., Wyder,S., Forslund,K., Heller,D., Huerta-Cepas,J., Simonovic,M., Roth,A., Santos,A., Tsafou,K.P., Kuhn,M., Bork,P., Jensen,L.J., and von Mering,C. (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.

The Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.

The UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204-12.

Watson,J.D. and Crick,F.H.C. (1953) Molecular structure of nucleic acids. *Nature*, **171**, 737–738.

Weaver,W. (1970) Molecular Biology: Origin of the Term. *Science*, **170**, 581–582.

Wetterstrand,K.A. (2016) DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program. *www.genome.gov/sequencingcostsdata*, Accessed [4 September 2016].

Whetzel,P.L., Noy,N.F., Shah,N.H., Alexander,P.R., Nyulas,C., Tudorache,T., and Musen,M.A. (2011) BioPortal: Enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res.*, **39**.

Wiener,N. (1949) Cybernetics or control and communication in the animal and the machine.

Wishart,D.S., Jewison,T., Guo,A.C., Wilson,M., Knox,C., Liu,Y., Djoumbou,Y., Mandal,R., Aziat,F., Dong,E., Bouatra,S., Sinelnikov,I., Arndt,D., Xia,J., Liu,P., Yallou,F., Bjorndahl,T.,

Perez-Pineiro,R., Eisner,R., Allen,F., Neveu,V., Greiner,R., and Scalbert,A. (2013) HMDB 3.0-The Human Metabolome Database in 2013. *Nucleic Acids Res.*, **41**.

Wu,M., Li,X., Kwoh,C.-K., and Ng,S.-K. (2009) A core-attachment based method to detect protein complexes in PPI networks. *BMC Bioinformatics*, **10**, 169.

Zhang,W. and Zou,X. (2015) A New Method for Detecting Protein Complexes based on the Three Node Cliques. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **12**, 879–86.

# 研究成果発表など

## 投稿論文・総説

- Shu Tadaka, Kengo Kinoshita, “NCMine: Core-peripheral based functional module detection using near-clique mining”, Bioinformatics, 2016
- Yuichi Aoki, Yasunobu Okamura, Shu Tadaka, Kengo Kinoshita, Takeshi Obayashi, “ATTED-II in 2016: A Plant Coexpression Database Towards Lineage-Specific Coexpression”, Plant Cell Physiology, Oxford Journals, Volume 57 Issue 1, pp.e5, 2016
- Yasunobu Okamura, Yuichi Aoki, Takeshi Obayashi, Shu Tadaka, Satoshi Ito, Takahumi Narise, Kengo Kinoshita, “COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems”, Nucleic Acids Research, Oxford Journals, Volume 43 (D1), pp.D82-D86, 2015
- Takeshi Obayashi, Yasunobu Okamura, Satoshi Ito, Shu Tadaka, Aoki Yuichi, Matsuyuki Shirota and Kengo Kinoshita. “ATTED-II in 2014: Evaluation of Gene Coexpression in Agriculturally Important Plants”, Plant Cell Physiology, Oxford Journals, Volume 55, pp.e6, 2014
- Katayama, T., Wilkinson, MD., Aoki-Kinoshita, KF., Kawashima, S., Yamamoto, Y., Yamaguchi, A., Okamoto, S., Kawano, S., Kim, JD., Wang, Y., Wu H., Kano, Y., Ono, H., Bono, H., Kocbek, S., Aerts, J., Akune, Y., Antezana, E., Arakawa, K., Aranda, B., Baran, J., Bolleman, J., Bonnal, RJ., Buttigieg PL., Campbell, MP., Chen, YA., Chiba, H., Cock, PJ., Cohen, KB., Constantin A., Duck, G., Dumontier, M., Fujisawa, T., Fujiwara, T., Goto, N., Hoehndorf R., Igarashi, Y., Itaya, H., Ito, M., Iwasaki, W., Kalas, M., Katoda, T., Kim T., Kokubu, A., Komiyama, Y., Kotera, M., Laibe, C., Lapp, H., Lütteke, T., Marshall, MS., Mori, T., Mori, H., Morita, M., Murakami, K., Nakao, M., Narimatsu, H., Nishide, H., Nishimura, Y., Nystrom-Persson, J., Ogishima S., Okamura, Y., Okuda, S., Oshita, K., Packer, NH., Prins, P., Ranzinger, R., RoccaSerra, P., Sansone, S., Sawaki, H., Shin, SH., Splendiani, A., Strozzi, F., Tadaka, S., Toukach, P., Uchiyama, I., Umezaki, M., Vos, R., Whetzel, PL., Yamada, I., Yamasaki, C., Yamashita, R., York, WS., Zmasek CM., Kawamoto, S. and Takagi, T., “BioHackathon series in 2011 and 2012: penetration of ontology and linked data in life science domains”, Journal of Biomed Semantics, 2014 Feb 5;5(1):5
- Takeshi Obayashi, Yasunobu Okamura, Satoshi Ito, Shu Tadaka, Ikuko IN Motoike and Kengo Kinsohita. “COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals”, Nucleic Acids Research, Oxford Journals, Volume 41 (D1), pp.D1014-D1020, 2013

# 国際学会における発表

## ポスター発表

- Shu Tadaka Yasunobu Okamura, Daisuke Saigusa, Ikuko N Motoike, Kengo Kinoshita, “MetaboloView: integrated analysis platform for large-scale metabolomics cohort datasets”, ISMB/ECCB 2015, 2015 年 7 月, Dublin (Ireland)
- Shu Tadaka,  Takeshi Obayashi, Kengo Kinoshita, “Detection of functional modules in protein networks by near-clique extraction”, GIW/ISCB-Asia, 2014 年 12 月, Tokyo

# 国内学会・研究会における発表

## 口頭発表

- 田高 周, 小館 俊, 上木 庸平, 小澤 和也, 佐藤 市也, 安澤 隼人, “オープンキャンパス向け案内支援・データ解析基盤 SmartCampus2 の開発”, 東北大学大学院情報科学研究科 平成 27 年度採択 学生プロジェクト 成果報告会, 2016 年 9 月, 東北大学（宮城）
- 田高 周, 小澤 和也, 南谷 和毅, 岡村 容伸, 小舘 俊, 安澤 隼人, 佐藤 壮真, 池野 直人, 大林 武, 木下 賢吾, "SmartCampus: オープンキャンパスに適した案内支援・データ解析基盤の開発", 日本情報処理学会 第 77 回全国大会, 2015 年 4 月, 京都大学（京都）
- Shu Tadaka, Takeshi Obayashi, Kengo Kinshita, “Identification of ufnctional modules in protein networks by near-clique detection algorithm”, 情報処理学会 バイオ情報学研究会 第 33 回バイオ情報学研究発表会, 2013 年 3 月, 東北大学（宮城）

## ポスター発表

- Shu Tadaka, Kengo Kinoshita, “Development and application of a simple yet flexible framework for construction of gene-pair relation database”, 生命医薬情報学連合大会, 2016 年 9 月, 国際交流会館プラザ平成（東京）
- Shu Tadaka, Kengo Kinoshita, “NCMine: Core-peripheral based functional module detection using near-clique mining”, The 3rd CWRU-Tohoku Joint workshop Collaboration on Data Science Engineering, 2016 年 8 月, 東北大学（宮城）
- 田高 周, 大林 武, 木下 賢吾, "NCMine: 相互作用ネットワーク内の機能モジュール検出のための新規手法の開発", 生命医薬情報学連合大会, 2014 年 10 月, 仙台国際センター（宮城）
- 田高 周, 大林 武, 木下 賢吾, "NCMine: 相互作用ネットワーク内のクラスタ構造を探るための新規クラスタリング手法の開発", 日本分子生物学会年会, 2013 年 12 月, 神戸国際展示場など（神戸）
- Shu Tadaka, Yasunobu Okamura, Takeshi Obayashi, Kengo Kinoshita, “ATTED-II meets the Semantic Web”, 生命医薬情報学連合大会, 2013 年 10 月, タワーホール船堀（東京）
- Shu Tadaka, Takeshi Obayashi, Kengo Kinoshita, “Identification of functional modules in protein networks by near-clique detection algorithm”, 生命医薬情報学連合大会, 2012 年 10 月, タワーホール船堀（東京）

# 受賞

- GIW/ISCB-Asia 2015 にて Poster Award 受賞

# プロジェクト採択

- 東北大学大学院情報科学研究科 平成 27 年度学生プロジェクト 採択
  - 研究題目:
    オープンキャンパス向け案内支援・データ解析基盤 SmartCampus2 の開発
  - 研究分担者:
    田高 周 1, 小館 俊 1, 上木 庸平 2, 小澤 和也 1, 佐藤 市也 2, 安澤 隼人 1 （1: 東北大学大学院情報科学研究科 木下・大林研究室, 2: 東北大学大学院情報科学研究科 篠原研究室）

# 謝辞

本研究を遂行し学位論文をまとめるにあたり、指導教官であった木下賢吾教授にお礼申し上げます。終始にわたり、数多くのご指導、ご助言、ご協力をいただきましたこと、深く感謝しております。また、大林武先生、西羽美先生をはじめとする、木下・大林研究室の皆様にも深くお礼申し上げます。木下・大林研究室は、木下教授を始めとする先生方、先輩・後輩に至るまで、様々な方とのディスカッションを通し有益な意見交換を行うことができる環境で、日々充実した研究室生活を送ることができました。

**謝辞**

# 付録

# 補足図表

表 13 図 9 に含まれる遺伝子一覧

| degree-membership (degree >= 150) | | | | | |
|---|---|---|---|---|---|
| GeneID | degree | Cluster membership | Pathway membership | Gene Symbol | Description |
| 7157 | 272 | 213 | 33 | TP53 | tumor protein p53 |
| 7532 | 249 | 126 | 5 | YWHAG | tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, gamma |
| 2033 | 211 | 173 | 24 | EP300 | E1A binding protein p300 |
| 6714 | 210 | 177 | 25 | SRC | SRC proto-oncogene, non-receptor tyrosine kinase |
| 1387 | 200 | 159 | 24 | CREBBP | CREB binding protein |
| 2885 | 195 | 151 | 34 | GRB2 | growth factor receptor-bound protein 2 |
| 2099 | 190 | 142 | 5 | ESR1 | estrogen receptor 1 |
| 4088 | 184 | 130 | 15 | SMAD3 | SMAD family member 3 |
| 5578 | 174 | 115 | 54 | PRKCA | protein kinase C, alpha |
| 1457 | 170 | 99 | 8 | CSNK2A1 | casein kinase 2, alpha 1 polypeptide |

| | | | | | |
|---|---|---|---|---|---|
| 4087 | 168 | 108 | 13 | SMAD2 | SMAD family member 2 |
| 1956 | 163 | 109 | 30 | EGFR | epidermal growth factor receptor |
| 5594 | 162 | 104 | 81 | MAPK1 | mitogen-activated protein kinase 1 |
| 6310 | 161 | 45 | 0 | ATXN1 | ataxin 1 |
| 7046 | 156 | 85 | 14 | TGFBR1 | transforming growth factor, beta receptor 1 |
| 4089 | 154 | 102 | 12 | SMAD4 | SMAD family member 4 |
| 2534 | 154 | 105 | 14 | FYN | FYN proto-oncogene, Src family tyrosine kinase |
| 367 | 152 | 112 | 2 | AR | androgen receptor |
| 56893 | 150 | 40 | 1 | UBQLN4 | ubiquilin 4 |
| degree-membership (membership >= 150) | | | | | |
| GeneID | degree | Cluster membership | Pathway membership | Gene Symbol | Description |
| 7157 | 272 | 213 | 33 | TP53 | tumor protein p53 |
| 6714 | 210 | 177 | 25 | SRC | SRC proto-oncogene, non-receptor tyrosine kinase |
| 2033 | 211 | 173 | 24 | EP300 | E1A binding protein p300 |
| 1387 | 200 | 159 | 24 | CREBBP | CREB binding protein |

| | | | | | |
|---|---|---|---|---|---|
| 2885 | 195 | 151 | 34 | GRB2 | growth factor receptor-bound protein 2 |
| 2099 | 190 | 142 | 5 | ESR1 | estrogen receptor 1 |
| 4088 | 184 | 130 | 15 | SMAD3 | SMAD family member 3 |
| 7532 | 249 | 126 | 5 | YWHAG | tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, gamma |
| 5578 | 174 | 115 | 54 | PRKCA | protein kinase C, alpha |
| 1956 | 163 | 109 | 30 | EGFR | epidermal growth factor receptor |
| 367 | 152 | 112 | 2 | AR | androgen receptor |
| 4087 | 168 | 108 | 13 | SMAD2 | SMAD family member 2 |
| 2534 | 154 | 105 | 14 | FYN | FYN proto-oncogene, Src family tyrosine kinase |
| 5295 | 130 | 105 | 68 | PIK3R1 | phosphoinositide-3-kinase, regulatory subunit 1 (alpha) |
| 4089 | 154 | 102 | 12 | SMAD4 | SMAD family member 4 |
| 6464 | 117 | 100 | 14 | SHC1 | SHC (Src homology 2 domain containing) transforming protein 1 |
| | | | | | |
| 3265 | 69 | 20 | 50 | HRAS | Harvey rat sarcoma viral oncogene homolog |
| 5582 | 48 | 5 | 49 | PRKCG | protein kinase C, gamma |
| 4893 | 13 | 1 | 48 | NRAS | neuroblastoma RAS viral (v-ras) |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | oncogene homolog |
| 4790 | 74 | 31 | 47 | NFKB1 | nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 |
| 5605 | 14 | 4 | 46 | MAP2K2 | mitogen-activated protein kinase kinase 2 |
| 7124 | 14 | 1 | 46 | TNF | tumor necrosis factor |
| 5599 | 70 | 32 | 42 | MAPK8 | mitogen-activated protein kinase 8 |
| 5601 | 39 | 8 | 42 | MAPK9 | mitogen-activated protein kinase 9 |
| 5602 | 15 | 1 | 41 | MAPK10 | mitogen-activated protein kinase 10 |

```
# Overall
procedure NCM
  input: graph G=(vertex V, edge E)
  output: Merged_Modules, Module_Parent_Map

  phase1:
    Modules = call NCM-FIND-MODULES (G)
  phase2:
    Merged_Modules, Module_Parent_Map = call NCM-MERGE-MODULES (G)

  return Merged_Modules, Module_Parent_Map
end procedure

# Phase 1
procedure NCM-FIND-MODULES
  input: graph G
  output: Modules

  calculate weight of each node in G

  Modules = {}
  for all v in G
    module = {v}

    # expand clusters
    for n in (neighbor nodes of module sorted by node weight):
        candidate_module = module + {n}



        if (cliqueness of candidate_module >= cliqueness threshold)
                or (cliqueness-change between module and candicate_module
                    <= cliqueness-change threshold) then
            module += n
        else
            break
        end if
    end for

    Modules += module
```

```
  end for

  return Modules
end procedure

# Phase 2
procedure NCM-MERGE-MODULES
  input: G, Modules (produced by Phase1: NCM-FIND-MODULES)
  output: Merged_Modules, Module_Parent_Map

  Merged_Modules = (empty set)
  Module_Parent_Map = (empty mapping)

  for all module in Modules
    overlapping_module = find highly overlapped module from Merged_Modules,
                         or null if overlapping module is not found
    if overlapping_module is not null then
        if cliqueness of union(module, overlapping_module)
                >= max(cliqueness of module, cliqueness of overlapping_module)
            # Case 1
            Merged_Modules = Merged_Modules + union(module, overlapping_module)
       else
            # Case 2
            Merged_Modules =
                Merged_Modules + intersection(module, overlapping_module) + module
            Module_Parent_Map[module, overlapping_module] =
                intersection(module, overlapping_module))
          end if
      else
        # Case 3
        Merged_Modules += module
      end if

      return Merged_Modules, Module_Parent_Mapping
  end for
end procedure
```

**図 45 NCMine アルゴリズム 疑似コード**

```
<rdf:RDF
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema"
    xmlns:atted="http://addte.jp/rdf/0.1/">

    <owl:Ontology rdf:about="http://atted.jp/rdf/atted-rdf-02.owl">
        <rdfs:label>ATTED-II ontology</rdfs:label>
        <rdfs:comment>The ATTED-II ontology.</rdfs:comment>
    </owl:Ontology>
    <owl:Class rdf:about="atted:gene_id">
        <rdfs:label>Gene ID</rdfs:label>
        <rdfs:comment>
               Gene ID represented in Bio2RDF's Entrez gene URI</rdfs:comment>
        <rdfs:domain rdfs:resource="atted:gene_pair"/>
    </owl:Class>
    <owl:Class rdf:about="atted:gene_pair">
        <rdfs:label>Pair of genes</rdfs:label>
        <rdfs:comment>Pair of two gene</rdfs:comment>
    </owl:Class>
    <owl:Class rdf:about="atted:gene_id_1">
        <rdfs:label>Gene ID 1</rdfs:label>
        <rdfs:comment>One of genes in a co-expressing gene pair</rdfs:comment>
        <rdfs:subClassOf rdfs:resource="atted:gene_id"/>
    </owl:Class>
    <owl:Class rdf:about="atted:gene_id_2">
        <rdfs:label>Gene ID 2</rdfs:label>
        <rdfs:comment>One of genes in a co-expressing gene pair</rdfs:comment>
        <rdfs:subClassOf rdfs:resource="atted:gene_id"/>
    </owl:Class>
```

```
    <owl:ObjectProperty rdf:about="atted:coexpression_measure">
        <rdfs:label>Coexpression measure</rdfs:label>
        <rdfs:comment>Measure for coexpression degree</rdfs:comment>
        <rdfs:domain rdf:resource="atted:gene_pair"/>
        <rdfs:range rdf:resource="xsd:float"/>
    </owl:ObjectProperty>
    <owl:ObjectProperty rdf:about="atted:mutual_rank">
        <rdfs:label>Mutual rank</rdfs:label>
        <rdfs:comment>The 'Mutual rank' coexpression measure</rdfs:comment>
        <rdfs:subPropertyOf rdfs:resource="atted:coexpression_measure"/>
    </owl:ObjectProperty>
    <owl:ObjectProperty rdf:about="atted:pcc">
        <rdfs:label>Pearson correlation</rdfs:label>
        <rdfs:comment>The 'Pearson correlation' coexpression measure</rdfs:comment>
        <rdfs:subPropertyOf rdfs:resource="atted:coexpression_measure"/>
    </owl:ObjectProperty>
</rdf:RDF>
```

**図 46 ATTED-II/COXPRESdb RDF データモデルスキーマ**